

TARGETED FUSED RIDGE ESTIMATION OF INVERSE COVARIANCE MATRICES FROM MULTIPLE HIGH-DIMENSIONAL DATA CLASSES

ANDERS ELLERN BILGRAU*, CAREL F.W. PEETERS*,
POUL SVANTE ERIKSEN, MARTIN BØGSTED, AND WESSEL N. VAN WIERINGEN†

ABSTRACT. We consider the problem of jointly estimating multiple precision matrices from (aggregated) high-dimensional data consisting of distinct classes. An ℓ_2 -penalized maximum-likelihood approach is employed. The suggested approach is flexible and generic, incorporating several other ℓ_2 -penalized estimators as special cases. In addition, the approach allows for the specification of target matrices through which prior knowledge may be incorporated and which can stabilize the estimation procedure in high-dimensional settings. The result is a targeted fused ridge estimator that is of use when the precision matrices of the constituent classes are believed to chiefly share the same structure while potentially differing in a number of locations of interest. It has many applications in (multi)factorial study designs. We focus on the graphical interpretation of precision matrices with the proposed estimator then serving as a basis for integrative or meta-analytic Gaussian graphical modeling. Situations are considered in which the classes are defined by data sets and/or (subtypes of) diseases. The performance of the proposed estimator in the graphical modeling setting is assessed through extensive simulation experiments. Its practical usability is illustrated by the differential network modeling of 11 large-scale diffuse large B-cell lymphoma gene expression data sets. The estimator and its related procedures are incorporated into the R-package `rags2ridges`.

Keywords: Differential network estimation; Gaussian graphical modeling; Generalized fused ridge; High-dimensional data; ℓ_2 -penalized maximum likelihood; Structural meta-analysis; Subclass data

1. INTRODUCTION

High-dimensional data are ubiquitous in modern statistics. Consequently, the fundamental problem of estimating the covariance matrix or its inverse (the precision matrix) has received renewed attention. Suppose we have n i.i.d. observations of a p -dimensional variate distributed as $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Gaussian log-likelihood parameterized in terms of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is then given by:

$$(1) \quad \mathcal{L}(\boldsymbol{\Omega}; \mathbf{S}) \propto \ln|\boldsymbol{\Omega}| - \text{tr}(\mathbf{S}\boldsymbol{\Omega}),$$

where \mathbf{S} is the sample covariance matrix. When $n > p$ the maximum of (1) is attained at the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\Omega}}^{\text{ML}} = \mathbf{S}^{-1}$. However, in the high-dimensional case, i.e., when $p > n$, the sample covariance matrix \mathbf{S} is singular and its inverse ceases to exist. Furthermore, when $p \approx n$, the sample covariance matrix may be ill-conditioned and the inversion becomes numerically unstable. Hence, these situations necessitate usage of regularization techniques.

Here, we study the simultaneous estimation of numerous precision matrices when multiple classes of high-dimensional data are present. Suppose \mathbf{y}_{ig} is a realization of a p -dimensional Gaussian random vector for $i = 1, \dots, n_g$ independent observations nested within $g = 1, \dots, G$ classes, each with class-dependent covariance $\boldsymbol{\Sigma}_g$, i.e., $\mathbf{y}_{ig} \sim \mathcal{N}_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ for each designated class g . Hence, for each class a data set consisting of the $n_g \times p$ matrix $\mathbf{Y}_g = [\mathbf{y}_{1g}, \dots, \mathbf{y}_{n_g g}]^\top$ is observed. Without loss of generality $\boldsymbol{\mu}_g = \mathbf{0}$ can be assumed as each data set \mathbf{Y}_g can be centered around its column means. The class-specific sample covariance matrix given by

$$\mathbf{S}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{y}_{ig} \mathbf{y}_{ig}^\top = \frac{1}{n_g} \mathbf{Y}_g^\top \mathbf{Y}_g,$$

* Shared first authorship.

† Principal corresponding author.

then constitutes the well-known MLE of Σ_g as discussed above. The closely related *pooled* sample covariance matrix

$$(2) \quad \mathbf{S}_\bullet = \frac{1}{n_\bullet} \sum_{g=1}^G \sum_{i=1}^{n_g} \mathbf{y}_{ig} \mathbf{y}_{ig}^\top = \frac{1}{n_\bullet} \sum_{g=1}^G n_g \mathbf{S}_g,$$

where $n_\bullet = \sum_{g=1}^G n_g$, is an oft-used estimate of the common covariance matrix across classes. In the high-dimensional case $p > n_\bullet$ (implying $p > n_g$) the \mathbf{S}_g and \mathbf{S}_\bullet are singular and their inverses do not exist. Our primary interest thus lies in estimating the precision matrices $\mathbf{\Omega}_1 = \mathbf{\Sigma}_1^{-1}, \dots, \mathbf{\Omega}_G = \mathbf{\Sigma}_G^{-1}$, as well as their commonalities and differences, when $p > n_\bullet$. We will develop a general ℓ_2 -penalized ML framework to this end which we designate *targeted fused ridge estimation*.

The estimation of multiple precision matrices from high-dimensional data classes is of interest in many applications. The field of oncogenomics, for example, often deals with high-dimensional data from high-throughput experiments. Class membership may have different connotations in such settings. It may refer to certain sub-classes within a single data set such as cancer subtypes (cancer is a very heterogeneous disease, even when present in a single organ). It may also designate different data sets or studies. Likewise, the class indicator may also refer to a conjunction of both subclass and study membership to form a two-way design of factors of interest (e.g., breast cancer subtypes present in a batch of study-specific data sets), as is often the case in oncogenomics. Our approach is thus motivated by the meta-analytic setting, where we aim for an integrative analysis in terms of simultaneously considering multiple data (sub-)classes, data sets, or both. Its desire is to borrow statistical power across classes by effectively increasing the sample size in order to improve sensitivity and specificity of discoveries.

1.1. Relation to literature and overview. There have been many proposals for estimating a single precision matrix in high-dimensional data settings. A popular approach is to amend (1) with an ℓ_1 -penalty [48, 2, 19, 49]. The solution to this penalized problem is generally referred to as the *graphical lasso* and it is popular as it performs automatic model selection, i.e., the resulting estimate is sparse. It is heavily used in Gaussian graphical modeling (GGM) as the support of a Gaussian precision matrix represents a Markov random field [24].

The ℓ_1 -approach has been extended to deal with more than a single sample-group. Guo et al. [21] have proposed a parametrization of class-specific precision matrices that expresses the individual elements as a product of shared and class-specific factors. They include ℓ_1 -penalties on both the shared and class-specific factors in order to jointly estimate the sparse precision matrices (representing graphical models). The penalty on the shared factors promotes a shared sparsity structure while the penalty on the class-specific factors promotes class-specific deviations from the shared sparsity structure. Danaher et al. [13] have generalized these efforts by proposing the *joint graphical lasso* which allows for various penalty structures. They study two particular choices: the *group graphical lasso* that encourages a shared sparsity structure across the class-specific precision matrices, and the *fused graphical lasso* that promotes a shared sparsity structure as well as shared precision element-values. A Bayesian approach to inferring multiple sparse precision matrices can be found in Peterson et al. [32].

While simultaneous estimation and model selection can be deemed elegant, automatic sparsity is not always an asset. It may be that one is intrinsically interested in more accurate representations of class-specific precision matrices for usage in, say, covariance-regularized regression [46] or discriminant analysis [33]. In such a situation one is not after sparse representations and one may prefer usage of a regularization method that shrinks the estimated elements of the precision matrices proportionally. In addition—when indeed considering network representations of data—the true class-specific graphical models need not be (extremely) sparse in terms of containing many zero elements. The ℓ_1 -penalty is unable to retrieve the sparsity pattern when the number of truly non-null elements exceeds the available sample size [42]. In such a situation one may wish to couple a non-sparsity-inducing penalty with a post-hoc selection step allowing for probabilistic control over element selection. We therefore consider ℓ_2 or ridge-type penalization.

In Section 2 the *targeted fused ridge estimation* framework will be presented. The proposed fused ℓ_2 -penalty allows for the simultaneous estimation of multiple precision matrices from high-dimensional data

classes that chiefly share the same structure but that may differentiate in locations of interest. The approach is targeted in the sense that it allows for the specification of target matrices that may encode prior information. The framework is flexible and general, containing the recent work of Price et al. [33] and van Wieringen and Peeters [42] as special cases. It may be viewed as a ℓ_2 -alternative to the work of Danaher et al. [13]. The method is contingent upon the selection of penalty values and target matrices, topics that are treated in Section 3. Section 4 then focuses on the graphical interpretation of precision matrices. It shows how the fused ridge precision estimates may be coupled with post-hoc support determination in order to arrive at multiple graphical models. We will refer to this coupling as the *fused graphical ridge*. This then serves as a basis for integrative or meta-analytic network modeling. Section 5 then assesses the performance of the proposed estimator through extensive simulation experiments. Section 6 illustrates the techniques by applying it in a large scale integrative study of gene expression data of diffuse large B-cell lymphoma. The focus is then on finding common motifs and motif differences in network representations of (deregulated) molecular pathways. Section 7 concludes with a discussion.

1.2. Notation. Some additional notation must be introduced. Throughout the text and supplementary material, we use the following notation for certain matrix properties and sets: We use $\mathbf{A} \succ \mathbf{0}$ and $\mathbf{B} \succeq \mathbf{0}$ to denote symmetric positive definite (p.d.) and positive semi-definite (p.s.d.) matrices \mathbf{A} and \mathbf{B} , respectively. By \mathbb{R} , \mathbb{R}_+ , and \mathbb{R}_{++} we denote the real numbers, the non-negative real numbers, and the strictly positive real numbers, respectively. In notational analogue, \mathcal{S}^p , \mathcal{S}_+^p , and \mathcal{S}_{++}^p are used to denote the space of $p \times p$ real symmetric matrices, the real symmetric p.s.d. matrices, and real symmetric p.d. matrices, respectively. That is, e.g., $\mathcal{S}_{++}^p = \{\mathbf{X} \in \mathbb{R}^{p \times p} : \mathbf{X} = \mathbf{X}^\top \wedge \mathbf{X} \succ \mathbf{0}\}$. Negative subscripts similarly denote negative reals and negative definiteness. By $\mathbf{A} \geq \mathbf{B}$ and similar we denote *element-wise* relations, i.e., $(\mathbf{A})_{jq} \geq (\mathbf{B})_{jq}$ for all (j, q) . Matrix subscripts will usually denote class membership, e.g., \mathbf{A}_g denotes (the realization of) matrix \mathbf{A} in class g . For notational brevity we will often use the shorthand $\{\mathbf{A}_g\}$ to denote the set $\{\mathbf{A}_g\}_{g=1}^G$.

The following notation is used throughout for operations: We write $\text{diag}(\mathbf{A})$ for the column vector composed of the diagonal of \mathbf{A} and $\text{vec}(\mathbf{A})$ for the vectorization operator which stacks the columns of \mathbf{A} on top of each other. Moreover, \circ will denote the Hadamard product while \otimes refers to the Kronecker product.

We will also repeatedly make use of several special matrices and functions. We let \mathbf{I}_p denote the $(p \times p)$ -dimensional identity matrix. Similarly, \mathbf{J}_p will denote the $(p \times p)$ -dimensional all-ones matrix. In addition, $\mathbf{0}$ will denote the null-matrix, the dimensions of which should be clear from the context. Lastly, $\|\cdot\|_F^2$ and $\mathbb{1}[\cdot]$ will stand for the squared Frobenius norm and the indicator function, respectively.

2. TARGETED FUSED RIDGE ESTIMATION

2.1. A general penalized log-likelihood problem. Suppose G classes of $(n_g \times p)$ -dimensional data exist and that the samples within each class are i.i.d. normally distributed. The log-likelihood for the data takes the following form under the additional assumption that all n_\bullet observations are independent:

$$(3) \quad \mathcal{L}(\{\mathbf{\Omega}_g\}; \{\mathbf{S}_g\}) \propto \sum_g n_g \{\ln|\mathbf{\Omega}_g| - \text{tr}(\mathbf{S}_g \mathbf{\Omega}_g)\}.$$

We desire to obtain estimates $\{\hat{\mathbf{\Omega}}_g\} \in \mathcal{S}_{++}^p$ of the precision matrices for each class. Though not a requirement, we primarily consider situations in which $p > n_g$ for all g , necessitating the need for regularization. To this end, amend (3) with the *fused ridge penalty* given by

$$(4) \quad f^{\text{FR}}(\{\mathbf{\Omega}_g\}; \{\lambda_{g_1 g_2}\}, \{\mathbf{T}_g\}) = \sum_g \frac{\lambda_{gg}}{2} \|\mathbf{\Omega}_g - \mathbf{T}_g\|_F^2 + \sum_{g_1, g_2} \frac{\lambda_{g_1 g_2}}{4} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2,$$

where the $\mathbf{T}_g \in \mathcal{S}_+^p$ indicate known class-specific *target matrices* (see also Section 3.3), the $\lambda_{gg} \in \mathbb{R}_{++}$ denote class-specific *ridge penalty parameters*, and the $\lambda_{g_1 g_2} \in \mathbb{R}_+$ are pair-specific *fusion penalty parameters* subject to the requirement that $\lambda_{g_1 g_2} = \lambda_{g_2 g_1}$. All penalties can then be conveniently summarized into a non-negative symmetric matrix $\mathbf{\Lambda} = [\lambda_{g_1 g_2}]$ which we call the *penalty matrix*. The diagonal of $\mathbf{\Lambda}$ corresponds to the class-specific ridge penalties whereas off-diagonal entries are the pair-specific fusion penalties. The

rationale and use of the penalty matrix is motivated further in Section 3.1. Combining (3) and (4) yields a general targeted fused ridge estimation problem:

$$(5) \quad \arg \max_{\{\boldsymbol{\Omega}_g\} \in \mathcal{S}_{++}^p} \left\{ \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}) - \sum_g \frac{\lambda_{gg}}{2} \|\boldsymbol{\Omega}_g - \mathbf{T}_g\|_F^2 - \sum_{g_1, g_2} \frac{\lambda_{g_1 g_2}}{4} \|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2 \right\}.$$

The problem of (5) is strictly concave. Furthermore, it is worth noting that non-zero fusion penalties, $\lambda_{g_1 g_2} > 0$ for all $g_1 \neq g_2$, alone will not guarantee uniqueness when $p > n_\bullet$: In high dimensions, all ridge penalties λ_{gg} should be strictly positive to ensure identifiability. These and other properties of the estimation problem are reviewed in Section 2.2.

The problem stated in (5) is very general. We shall sometimes consider a single common ridge penalty $\lambda_{gg} = \lambda$ for all g , as well as a common fusion penalty $\lambda_{g_1 g_2} = \lambda_f$ for all class pairs $g_1 \neq g_2$ (cf., however, Section 3.1) such that $\boldsymbol{\Lambda} = \lambda \mathbf{I}_p + \lambda_f (\mathbf{J}_p - \mathbf{I}_p)$. This simplification leads to the first special case:

$$\arg \max_{\{\boldsymbol{\Omega}_g\} \in \mathcal{S}_{++}^p} \left\{ \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}) - \frac{\lambda}{2} \sum_g \|\boldsymbol{\Omega}_g - \mathbf{T}_g\|_F^2 - \frac{\lambda_f}{4} \sum_{g_1, g_2} \|(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2 \right\}.$$

Here and analogous to (5), λ controls the rate of shrinkage of each precision $\boldsymbol{\Omega}_g$ towards the corresponding target \mathbf{T}_g [42], while λ_f determines the retainment of entry-wise similarities between $(\boldsymbol{\Omega}_{g_1} - \mathbf{T}_{g_1})$ and $(\boldsymbol{\Omega}_{g_2} - \mathbf{T}_{g_2})$ for all class pairs $g_1 \neq g_2$.

When $\mathbf{T}_g = \mathbf{T}$ for all g , the problem further simplifies to

$$(6) \quad \arg \max_{\{\boldsymbol{\Omega}_g\} \in \mathcal{S}_{++}^p} \left\{ \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}) - \frac{\lambda}{2} \sum_g \|\boldsymbol{\Omega}_g - \mathbf{T}\|_F^2 - \frac{\lambda_f}{4} \sum_{g_1, g_2} \|\boldsymbol{\Omega}_{g_1} - \boldsymbol{\Omega}_{g_2}\|_F^2 \right\},$$

where the targets are seen to disappear from the fusion term. Lastly, when $\mathbf{T} = \mathbf{0}$ the problem (6) reduces to its simplest form recently considered by Price et al. [33]. Appendix A studies, in order to support an intuitive feel for the fused ridge estimation problem, its geometric interpretation in this latter context.

2.2. Estimator and properties. There is no explicit solution to (5) except for certain special cases and thus an iterative optimization procedure is needed for its general solution. As described in Section 2.3, we employ a coordinate ascent procedure which relies on the concavity of the penalized likelihood (see Lemma 3 in Appendix B.1) and repeated use of the following result, whose proof (as indeed all proofs) has been deferred to Appendix B.2:

Proposition 1. *Let $\{\mathbf{T}_g\} \in \mathcal{S}_{++}^p$ and let $\boldsymbol{\Lambda} \in \mathcal{S}^G$ be a fixed penalty matrix such that $\boldsymbol{\Lambda} \geq \mathbf{0}$ and $\text{diag}(\boldsymbol{\Lambda}) > \mathbf{0}$. Furthermore, assume that $\boldsymbol{\Omega}_g$ is p.d. and fixed for all $g \neq g_0$. The maximizing argument for class g_0 of the optimization problem (5) is then given by*

$$(7) \quad \hat{\boldsymbol{\Omega}}_{g_0}(\boldsymbol{\Lambda}, \{\boldsymbol{\Omega}_g\}_{g \neq g_0}) = \left\{ \left[\bar{\lambda}_{g_0} \mathbf{I}_p + \frac{1}{4} (\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} \bar{\mathbf{T}}_{g_0})^2 \right]^{1/2} + \frac{1}{2} (\bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} \bar{\mathbf{T}}_{g_0}) \right\}^{-1},$$

where

$$(8) \quad \bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} - \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}} (\boldsymbol{\Omega}_g - \mathbf{T}_g), \quad \bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0}, \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0 \bullet}}{n_{g_0}},$$

with $\lambda_{g_0 \bullet} = \sum_g \lambda_{gg_0}$ denoting the sum of the g_0 th column (or row) of $\boldsymbol{\Lambda}$.

Remark 1. Defining $\bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0}$ in Proposition 1 may be deemed redundant. However, it allows us to state equivalent alternatives to (8) without confusing notation. See Section 2.3 as well as Appendix B.2 and Section 1 of the Supplementary Material.

Remark 2. The target matrices from Proposition 1 may be chosen nonnegative definite. However, choosing n.d. targets may lead to ill-conditioned estimates in the limit. From a shrinkage perspective we thus prefer to choose $\{\mathbf{T}_g\} \in \mathcal{S}_{++}^p$. See Section 3.3.

Proposition 1 provides a function for updating the estimate of the g_0 th class while fixing the remaining parameters. As a special case, consider the following. If all off-diagonal elements of $\mathbf{\Lambda}$ are zero no ‘class fusion’ of the estimates takes place and the maximization problem decouples into G individual, disjoint ridge estimations: See Corollary 1 in Appendix B.2. The next result summarizes some properties of (7):

Proposition 2. *Consider the estimator of Proposition 1 and its accompanying assumptions. Let $\hat{\mathbf{\Omega}}_g \equiv \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g})$ be the precision matrix estimate of the g th class. For this estimator, the following properties hold:*

- i. $\hat{\mathbf{\Omega}}_g \succ \mathbf{0}$ for all $\lambda_{gg} \in \mathbb{R}_{++}$;
- ii. $\lim_{\lambda_{gg} \rightarrow 0^+} \hat{\mathbf{\Omega}}_g = \mathbf{S}_g^{-1}$ if $\sum_{g' \neq g} \lambda_{gg'} = 0$ and $p \leq n_g$;
- iii. $\lim_{\lambda_{gg} \rightarrow \infty} \hat{\mathbf{\Omega}}_g = \mathbf{T}_g$ if $\lambda_{gg'} < \infty$ for all $g' \neq g$;
- iv. $\lim_{\lambda_{g_1 g_2} \rightarrow \infty} (\hat{\mathbf{\Omega}}_{g_1} - \mathbf{T}_{g_1}) = \lim_{\lambda_{g_1 g_2} \rightarrow \infty} (\hat{\mathbf{\Omega}}_{g_2} - \mathbf{T}_{g_2})$ if $\lambda_{g'_1 g'_2} < \infty$ for all $\{g'_1, g'_2\} \neq \{g_1, g_2\}$.

The first item of Proposition 2 implies that strictly positive λ_{gg} are sufficient to guarantee p.d. estimates from the ridge estimator. The second item then implies that if ‘class fusion’ is absent one obtains as the right-hand limit for group g the standard MLE \mathbf{S}_g^{-1} , whose existence is only guaranteed when $p \leq n_g$. The third item shows that the fused ridge precision estimator for class g is shrunk exactly to its target matrix when the ridge penalty tends to infinity while the fusion penalties do not. The last item shows that the precision estimators of any two classes tend to a common estimate when the fusion penalty between them tends to infinity while all remaining penalty parameters remain finite.

The attractiveness of the general estimator hinges upon the efficiency by which it can be obtained. We state a result useful in this respect before turning to our computational approach in Section 2.3:

Proposition 3. *Let $\hat{\mathbf{\Omega}}_g \equiv \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g})$ be the precision matrix estimate (7) for the g th class and define $[\hat{\mathbf{\Omega}}_g]^{-1} \equiv \hat{\mathbf{\Sigma}}_g$. The estimate $\hat{\mathbf{\Omega}}_g$ can then be obtained without inversion through:*

$$\hat{\mathbf{\Omega}}_g = \frac{1}{\bar{\lambda}_g} \left[\hat{\mathbf{\Sigma}}_g - (\bar{\mathbf{S}}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g) \right] = \frac{1}{\bar{\lambda}_g} \left\{ \left[\bar{\lambda}_g \mathbf{I}_p + \frac{1}{4} (\bar{\mathbf{S}}_g - \bar{\lambda}_g \bar{\mathbf{T}}_{g_0})^2 \right]^{1/2} - \frac{1}{2} (\bar{\mathbf{S}}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g) \right\}.$$

2.3. Algorithm. Equation (7) allows for updating the precision estimate $\hat{\mathbf{\Omega}}_g$ of class g by plugging in the remaining $\hat{\mathbf{\Omega}}_{g'}$, $g' \neq g$, and assuming them fixed. Hence, from initial estimates, all precision estimates may be iteratively updated until some convergence criterion is reached. We propose a block coordinate ascent procedure to solve (5) by repeated use of the results in Proposition 1. This procedure is outlined in Algorithm 1. By the strict concavity of the problem in (5), the procedure guarantees that, contingent upon convergence, the unique maximizer is attained when considering all $\hat{\mathbf{\Omega}}_g$ jointly. Moreover, we can state the following result:

Proposition 4. *The gradient ascent procedure given in Algorithm 1 will always stay within the realm of positive definite matrices \mathcal{S}_{++}^p .*

The procedure is implemented in the **rags2ridges** package within the R statistical language [34]. This implementation focuses on *stability* and *efficiency*. With regard to the former: Equivalent (in terms of the obtained estimator) alternatives to (8) can be derived that are numerically more stable for extreme values of $\mathbf{\Lambda}$. The most apparent such alternative is:

$$(9) \quad \bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0}, \quad \bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0} + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{\lambda_{g_0 \bullet}} (\mathbf{\Omega}_g - \mathbf{T}_g), \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0 \bullet}}{n_{g_0}}.$$

It ‘updates’ the target $\bar{\mathbf{T}}_g$ instead of the covariance $\bar{\mathbf{S}}_g$ and has the intuitive interpretation that the target matrix for a given class in the fused case is a combination of the actual class target matrix and the ‘target corrected’ estimates of remaining classes. The implementation makes use of this alternative where appropriate. See Section 1 of the Supplementary Material for details on alternative updating schemes.

Algorithm 1 Pseudocode for the fused ridge block coordinate ascent procedure.

```

1: Input:
2: Sufficient data:  $(\mathbf{S}_1, n_1), \dots, (\mathbf{S}_G, n_G)$ 
3: Penalty matrix:  $\mathbf{\Lambda}$ 
4: Convergence criterion:  $\varepsilon > 0$ 
5: Output:
6: Estimates:  $\hat{\mathbf{\Omega}}_1, \dots, \hat{\mathbf{\Omega}}_G$ 
7: procedure RIDGEP.FUSED( $\mathbf{S}_1, \dots, \mathbf{S}_G, n_1, \dots, n_G, \mathbf{\Lambda}, \varepsilon$ )
8:   Initialize:  $\hat{\mathbf{\Omega}}_g^{(0)}$  for all  $g$ .
9:   for  $c = 1, 2, 3, \dots$  do
10:    for  $g = 1, 2, \dots, G$  do
11:      Update  $\hat{\mathbf{\Omega}}_g^{(c)} := \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \hat{\mathbf{\Omega}}_1^{(c)}, \dots, \hat{\mathbf{\Omega}}_{g-1}^{(c)}, \hat{\mathbf{\Omega}}_{g+1}^{(c-1)}, \dots, \hat{\mathbf{\Omega}}_G^{(c-1)})$  by (7).
12:    end for
13:    if  $\max_g \left\{ \frac{\|\hat{\mathbf{\Omega}}_g^{(c)} - \hat{\mathbf{\Omega}}_g^{(c-1)}\|_F^2}{\|\hat{\mathbf{\Omega}}_g^{(c)}\|_F^2} \right\} < \varepsilon$  then
14:      return  $(\hat{\mathbf{\Omega}}_1^{(c)}, \dots, \hat{\mathbf{\Omega}}_G^{(c)})$ 
15:    end if
16:  end for
17: end procedure

```

Efficiency is secured through various roads. First, in certain special cases closed-form solutions to (5) exist. When appropriate, these explicit solutions are used. Moreover, these solutions may provide warm-starts for the general problem. See Section 2 of the Supplementary Material for details on estimation in these special cases. Second, the result from Proposition 3 is used, meaning that the relatively expensive operation of matrix inversion is avoided. Third, additional computational speed was achieved by implementing core operations in C++ via the R-packages `Rcpp` and `RcppArmadillo` [15, 16, 18, 38]. These efforts make analyzes with large p feasible. Throughout, we will initialize the algorithm with $\hat{\mathbf{\Omega}}_g^{(0)} = p / \text{tr}(\mathbf{S}_\bullet) \cdot \mathbf{I}_p$ for all g .

3. PENALTY AND TARGET SELECTION

3.1. The penalty graph and analysis of factorial designs. Equality of all class-specific ridge penalties λ_{gg} is deemed restrictive, as is equality of all pair-specific fusion penalties $\lambda_{g_1 g_2}$. In many settings, such as the analysis of factorial designs, finer control over the individual values of λ_{gg} and $\lambda_{g_1 g_2}$ befits the analysis. This will be motivated by several examples of increasing complexity. In order to do so, some additional notation is developed: The penalties of $\mathbf{\Lambda}$ can be summarized by a node- and edge-weighted graph $\mathcal{P} = (W, H)$ where the vertex set W corresponds to the possible classes and the edge set H corresponds to the similarities to be retained. The weight of node $g \in W$ is given by λ_{gg} and the weight of edge $(g_1, g_2) \in H$ is then given by $\lambda_{g_1 g_2}$. We refer to \mathcal{P} as the *penalty graph* associated with the penalty matrix $\mathbf{\Lambda}$. The penalty graph \mathcal{P} is simple and undirected as the penalty matrix is symmetric.

Example 1. Consider $G = 2$ classes or subtypes (ST) of diffuse large B-cell lymphoma (DLBCL) patients with tumors resembling either so-called activated B-cells (ABC) or germinal centre B-cells (GCB). Patients with the latter subtype have superior overall survival [1]. As the GCB phenotype is more common than ABC, one might imagine a scenario where the two class sample sizes are sufficiently different such that $n_{\text{GCB}} \gg n_{\text{ABC}}$. Numeric procedures to obtain a common ridge penalty (see, e.g., Section 3.2) would then be dominated by the smaller group. Hence, choosing non-equal class ridge penalties for each group will allow for a better analysis. In such a case, the following penalty graph and matrix would be suitable:

$$(10) \quad \mathcal{P} = \begin{array}{ccc} & \text{ABC} & \text{GCB} \\ & \circ & \circ \\ \lambda_{11} & \text{---} \lambda_f & \lambda_{22} \end{array} \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_f \\ \lambda_f & \lambda_{22} \end{bmatrix}.$$

Example 2. Consider data from a one-way factorial design where the factor is ordinal with classes A, B, and C. For simplicity, we choose the same ridge penalty λ for each class. Say we have prior information that A is closer to B and B is closer to C than A is to C. The fusion penalty on the pairs containing the intermediate level B might then be allowed to be stronger. The following penalty graph and matrix are thus sensible:

$$(11) \quad \mathcal{P} = \begin{array}{ccccc} & \text{A} & & \text{B} & & \text{C} \\ & \circ & & \circ & & \circ \\ & \lambda & & \lambda & & \lambda \\ & \diagdown & & \diagup & & \diagdown \\ & & \lambda_{AC} & & & \end{array} \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda & \lambda_B & \lambda_{AC} \\ \lambda_B & \lambda & \lambda_B \\ \lambda_{AC} & \lambda_B & \lambda \end{bmatrix}.$$

Depending on the application, one might even omit the direct shrinkage between A and C by fixing $\lambda_{AC} = 0$. A similar penalty scheme might also be relevant if one class of the factor is an unknown mix of the remaining classes and one wishes to borrow statistical power from such a class.

Example 3. In two-way or n -way factorial designs one might wish to retain similarities in the ‘direction’ of each factor along with a factor-specific penalty. Consider, say, 3 oncogenic data sets (DS_1 , DS_2 , DS_3) regarding ABC and GCB DLBCL cancer patients. This yields a total of $G = 6$ classes of data. One choice of penalization of this 2 by 3 design is represented by the penalty graph and matrix below:

$$(12) \quad \mathcal{P} = \begin{array}{ccccc} & \text{DS}_1 & & \text{DS}_2 & & \text{DS}_3 \\ & \circ & & \circ & & \circ \\ \text{GCB} & \lambda & & \lambda & & \lambda \\ & \diagdown & & \diagup & & \diagdown \\ & & \lambda_{ST} & & & \\ \text{ABC} & \circ & & \circ & & \circ \\ & \diagup & & \diagdown & & \diagup \\ & & \lambda_{ST} & & & \end{array} \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda & \lambda_{DS} & \lambda_{DS} & \lambda_{ST} & 0 & 0 \\ \lambda_{DS} & \lambda & \lambda_{DS} & 0 & \lambda_{ST} & 0 \\ \lambda_{DS} & \lambda_{DS} & \lambda & 0 & 0 & \lambda_{ST} \\ \lambda_{ST} & 0 & 0 & \lambda & \lambda_{DS} & \lambda_{DS} \\ 0 & \lambda_{ST} & 0 & \lambda_{DS} & \lambda & \lambda_{DS} \\ 0 & 0 & \lambda_{ST} & \lambda_{DS} & \lambda_{DS} & \lambda \end{bmatrix}.$$

This example would favor similarities (with the same force) only between pairs sharing a common level in each factor. This finer control allows users, or the employed algorithm, to penalize differences between data sets more (or less) strongly than differences between the ABC and GCB sub-classes. This corresponds to not applying direct shrinkage of interaction effects which is of interest in some situations.

While the penalty graph primarily serves as an intuitive overview, it does provide some aid in the construction of the penalty matrix for multifactorial designs. For example, the construction of the penalty matrix (12) in Example 3 corresponds to a Cartesian graph product of two complete graphs similar to those given in (10) and (11). We state that \mathcal{P} and $\mathbf{\Lambda}$ should be chosen carefully in conjunction with the choice of target matrices. Ideally, only strictly necessary penalization parameters (from the perspective of the desired analysis) should be introduced. Each additional penalty introduced will increase the difficulty of finding the optimal penalty values by increasing the dimension of the search-space.

3.2. Selection of penalty parameters. As the ℓ_2 -penalty does not automatically induce sparsity in the estimate, it is natural to seek loss efficiency. We then use cross-validation (CV) for penalty parameter selection due to its relation to the minimization of the Kullback-Leibler divergence and its predictive accuracy stemming from its data-driven nature. Randomly divide the data of each class into $k = 1, \dots, K$ disjoint subsets of approximately the same size. Previously, we have defined $\hat{\Omega}_g \equiv \hat{\Omega}_g(\mathbf{\Lambda}, \{\Omega_{g'}\}_{g' \neq g})$ to be the precision matrix estimate of the g th class. Let $\hat{\Omega}_g^{-k}$ be the analogous estimate (with similar notational dependencies) for class g based on all samples not in k . Also, let \mathbf{S}_g^k denote the sample covariance matrix for class g based on the data in subset k and let n_g^k denote the size of subset k in class g . The K -fold CV score for our fused regularized precision estimate based on the fixed penalty $\mathbf{\Lambda}$ can then be given as:

$$\text{KCV}(\mathbf{\Lambda}) = \frac{1}{KG} \sum_{g=1}^G \sum_{k=1}^K n_g^k \left[-\ln |\hat{\Omega}_g^{-k}| + \text{tr}(\hat{\Omega}_g^{-k} \mathbf{S}_g^k) \right] = -\frac{1}{KG} \sum_{g=1}^G \sum_{k=1}^K \mathcal{L}_g^k(\hat{\Omega}_g^{-k}; \mathbf{S}_g^k).$$

One would then choose $\mathbf{\Lambda}^*$ such that

$$(13) \quad \mathbf{\Lambda}^* = \arg \min_{\mathbf{\Lambda}} \text{KCV}(\mathbf{\Lambda}), \quad \text{subject to: } \mathbf{\Lambda} \geq \mathbf{0} \wedge \text{diag}(\mathbf{\Lambda}) > \mathbf{0}.$$

The least biased predictive accuracy can be obtained by choosing $K = n_g$ such that $n_g^k = 1$. This would give the fused version of leave-one-out CV (LOOCV). Unfortunately, LOOCV is computationally demanding for large p and/or large n_g . We propose to select the penalties by the computationally expensive LOOCV only if adequate computational power is available. In cases where it is not, we propose two alternatives.

Our first alternative is a special version of the LOOCV scheme that significantly reduces the computational cost. The *special* LOOCV (SLOOCV) is computed much like the LOOCV. However, only the class estimate in the class of the omitted datum is updated. More specifically, the SLOOCV problem is given by:

$$(14) \quad \mathbf{\Lambda}^\diamond = \arg \min_{\mathbf{\Lambda}} \text{SLOOCV}(\mathbf{\Lambda}), \quad \text{subject to: } \mathbf{\Lambda} \geq \mathbf{0} \wedge \text{diag}(\mathbf{\Lambda}) > \mathbf{0},$$

with

$$\text{SLOOCV}(\mathbf{\Lambda}) = -\frac{1}{n_\bullet} \sum_{g=1}^G \sum_{i=1}^{n_g} \mathcal{L}_g^i(\tilde{\mathbf{\Omega}}_g^{-i}; \mathbf{S}_g^i).$$

The estimate $\tilde{\mathbf{\Omega}}_g^{-i}$ in (14) is obtained by updating only $\hat{\mathbf{\Omega}}_g$ using Proposition 1. For all other $g' \neq g$, $\tilde{\mathbf{\Omega}}_g^{-i} = \hat{\mathbf{\Omega}}_{g'}$. The motivation for the SLOOCV is that a single observation in a given class g does not exert heavy direct influence on the estimates in the other classes. This way the number of fused ridge estimations for each given $\mathbf{\Lambda}$ and each given leave-one-out sample is reduced from n_\bullet to G estimations. Our second and fastest alternative is an approximation of the fused LOOCV score. This approximation can be used as an alternative to (S)LOOCV when the class sample sizes are relatively large (precisely the scenario where LOOCV is unfeasible). See Section 3 of the Supplementary Material for detailed information on this approximation.

3.3. Choice of target matrices. The target matrices $\{\mathbf{T}_g\}$ can be used to encode prior information and their choice is highly dependent on the application at hand. As they influence the efficacy as well as the amount of bias of the estimate, it is of some importance to make a well-informed choice. Here, we describe several options of increasing level of informativeness.

In the non-fused setting, the consideration of a scalar target matrix $\mathbf{T} = \alpha \mathbf{I}_p$ for some $\alpha \in [0, \infty)$ leads to a computational benefit stemming from the property of rotation equivariance [42]: Under such targets the ridge estimator only operates on the eigenvalues of the sample covariance matrix. This benefit transfers to the fused setting for the estimator described in Proposition 1. Hence, one may consider $\mathbf{T}_g = \alpha_g \mathbf{I}_p$ with $\alpha_g \in [0, \infty)$ for each g . The limited fused ridge problem in Price et al. [33] corresponds to choosing $\alpha_g = 0$ for all g , such that a common target $\mathbf{T}_g = \mathbf{T} = \mathbf{0}$ is employed. This can be considered the least informative target possible. We generally argue against the use of the non p.d. target $\mathbf{T} = \mathbf{0}$, as it implies shrinking the class precision matrices towards the null matrix and thus towards infinite variance.

Choosing α_g to be strictly positive implies a (slightly) more informative choice. The rotation equivariance property dictates that it is sensible to choose α_g based on empirical information regarding the eigenvalues of \mathbf{S}_g . One such choice could be the average of the reciprocals of the non-zero eigenvalues of \mathbf{S}_g . A straightforward alternative would be to choose $\alpha_g = [\text{tr}(\mathbf{S}_g)/p]^{-1}$. In the special case of (6) where all $\alpha_g = \alpha$ the analogous choice would be $\alpha = [\text{tr}(\mathbf{S}_\bullet)/p]^{-1}$.

More informative targets would move beyond the scalar matrix. An example would be the consideration of factor-specific targets for factorial designs. Recalling Example 3, one might deem the data set factor to be a ‘nuisance factor’. Hence, one might choose different targets \mathbf{T}_{GCB} and \mathbf{T}_{ABC} based on training data or the pooled estimates of the GCB and ABC samples, respectively. In general, the usage of pilot training data or (pathway) database information (or both) allows for the construction of target matrices with higher specificity. We illustrate how to construct targets from database information in the DLBCL application of Section 6.

4. FUSED GRAPHICAL MODELING

4.1. To fuse or not to fuse. As a preliminary step to downstream modeling one might consider testing the hypothesis of no class heterogeneity—and therefore the necessity of fusing—amongst the class-specific precision matrices. Effectively, one then wishes to test the null-hypothesis $H_0 : \boldsymbol{\Omega}_1 = \dots = \boldsymbol{\Omega}_G$. Under H_0 an explicit estimator is available in which the fused penalty parameters play no role, cf. Section 2.2 of the Supplementary Material. Here we suggest a score test [6] for the evaluation of H_0 in conjunction with a way to generate its null distribution in order to assess its observational extremity.

A score test is convenient as it only requires estimation under the null hypothesis, allowing us to exploit the availability of an explicit estimator. The score statistic equals:

$$U = - \sum_{g=1}^G \left(\frac{\partial \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\})}{\partial \boldsymbol{\Omega}_g} \right)^\top \left(\frac{\partial^2 \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\})}{\partial \boldsymbol{\Omega}_g \partial \boldsymbol{\Omega}_g^\top} \right)^{-1} \frac{\partial \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\})}{\partial \boldsymbol{\Omega}_g} \bigg|_{\boldsymbol{\Omega}_g = \hat{\boldsymbol{\Omega}}^{H_0}},$$

where $\hat{\boldsymbol{\Omega}}^{H_0}$ denotes the precision estimate under H_0 given in (S4), which holds for all classes g . The gradient can be considered in vectorized form and is readily available from (25). The Hessian of the log-likelihood equals $\partial^2 \mathcal{L} / (\partial \boldsymbol{\Omega}_g \partial \boldsymbol{\Omega}_g^\top) = -\boldsymbol{\Omega}_g^{-1} \otimes \boldsymbol{\Omega}_g^{-1}$. For practical purposes of evaluating the score statistic, we employ the identity $(\mathbf{A}^\top \otimes \mathbf{B}) \text{vec}(\mathbf{C}) = \text{vec}(\mathbf{BCA})$ which avoids the manipulation of $(p^2 \times p^2)$ -dimensional matrices. Hence, the test statistic U is computed by

$$\hat{U} = \sum_{g=1}^G \text{vec}(\hat{\mathbf{X}}_g)^\top \text{vec}(\hat{\boldsymbol{\Omega}}^{H_0} \hat{\mathbf{X}}_g \hat{\boldsymbol{\Omega}}^{H_0}) = \sum_{g=1}^G \text{tr}[\hat{\mathbf{X}}_g (\hat{\boldsymbol{\Omega}}^{H_0} \hat{\mathbf{X}}_g \hat{\boldsymbol{\Omega}}^{H_0})],$$

where $\hat{\mathbf{X}}_g = n_g \{2[(\hat{\boldsymbol{\Omega}}^{H_0})^{-1} - \mathbf{S}_g] - [(\hat{\boldsymbol{\Omega}}^{H_0})^{-1} - \mathbf{S}_g] \circ \mathbf{I}_p\}$.

The null distribution of U can be generated by permutation of the class labels: one permutes the class labels, followed by re-estimation of $\boldsymbol{\Omega}$ under H_0 and the re-calculation of the test statistic. The observed test statistic (under H_0) \hat{U} is obtained from the non-permuted class labels and the regular fused estimator. The p -value is readily obtained by comparing the observed test statistic \hat{U} to the null distribution obtained from the test statistic under permuted class labels. We note that the test is conditional on the choice of λ_{gg} .

4.2. Graphical modeling. A contemporary use for precision matrices is found in the reconstruction and analysis of networks through graphical modeling. Graphical models merge probability distributions of random vectors with graphs that express the conditional (in)dependencies between the constituent random variables. In the fusion setting one might think that the class precisions share a (partly) common origin (conditional independence graph) to which fusion appeals. We focus on class-specific graphs $\mathcal{G}_g = (V, E_g)$ with a finite set of vertices (or nodes) V and set of edges E_g . The vertices correspond to a collection of random variables and we consider the same set $V = \{Y_1, \dots, Y_p\}$ of cardinality p for all classes g . That is, we consider the same p variables in all G classes. The edge set E_g is a collection of pairs of distinct vertices $(Y_j, Y_{j'})$ that are connected by an undirected edge and this collection may differ between classes. In case we assume $\{Y_1, \dots, Y_p\} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_g)$ for all classes g we are considering multiple Gaussian graphical models.

Conditional independence between a pair of variables in the Gaussian graphical model corresponds to zero entries in the (class-specific) precision matrix. Let $\hat{\boldsymbol{\Omega}}_g$ denote a generic estimate of the precision matrix in class g . Then the following relations hold for all pairs $\{Y_j, Y_{j'}\} \in \mathcal{V}$ with $j \neq j'$:

$$(\hat{\boldsymbol{\Omega}}_g)_{jj'} = \omega_{jj'}^{(g)} = 0 \iff Y_j \perp\!\!\!\perp Y_{j'} \mid V \setminus \{Y_j, Y_{j'}\} \text{ in class } g \iff (Y_j, Y_{j'}) \notin E_g.$$

Hence, determining the (in)dependence structure of the variables for class g —or equivalently the edge set E_g of \mathcal{G}_g —amounts to determining the support of $\hat{\boldsymbol{\Omega}}_g$.

4.3. Edge selection. We stress that support determination may be skipped entirely as the estimated precision matrices can be interpreted as complete (weighted) graphs. For more sparse graphical representations we resort to support determination by a local false discovery rate (IFDR) procedure [17] proposed by Schäfer and

Strimmer [39]. This procedure assumes that the nonredundant off-diagonal entries of the partial correlation matrix

$$(\hat{\mathbf{P}}_g)_{jj'} = -\hat{\omega}_{jj'}^{(g)} \left(\hat{\omega}_{jj}^{(g)} \hat{\omega}_{j'j'}^{(g)} \right)^{-\frac{1}{2}}$$

follow a mixture distribution representing null and present edges. The null-distribution is known to be a scaled beta-distribution which allows for estimating the IFDR:

$$\widehat{\text{IFDR}}_{jj'}^{(g)} = P\left((Y_j, Y_{j'}) \notin E_g \mid (\hat{\mathbf{P}}_g)_{jj'}\right),$$

which gives the empirical posterior probability that the edge between Y_j and $Y_{j'}$ is null in class g conditional on the observed corresponding partial correlation. The analogous probability that an edge is present can be obtained by considering $1 - \widehat{\text{IFDR}}_{jj'}^{(g)}$. See [17, 39, 42] for further details on the IFDR procedure. Our strategy will be to select for each class only those edges for which $1 - \widehat{\text{IFDR}}_{jj'}^{(g)}$ surpasses a certain threshold (see Section 6). This two-step procedure of regularization followed by subsequent support determination has the advantage that it enables probabilistic statements about the inclusion (or exclusion) of edges.

4.4. Common and differential (sub-)networks. After estimation and sparsification of the class precision matrices the identification of commonalities and differences between the graphical estimates are of natural interest. Here we consider some (summary) measures to aid such identifications. Assume in the following that multiple graphical models have been identified by the sparsified estimates $\hat{\Omega}_1^0, \dots, \hat{\Omega}_G^0$ and that the corresponding graphs are denoted by $\mathcal{G}_1, \dots, \mathcal{G}_G$.

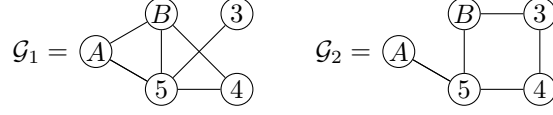
An obvious method of comparison is by pairwise graph differences or intersections. We utilize the *differential network* $\mathcal{G}_{g_1 \setminus g_2} = (V, E_{g_1} \setminus E_{g_2})$ between class g_1 and g_2 to provide an overview of edges present in one class but not the other. The *common network* $\mathcal{G}_{1 \cap 2} = (V, E_1 \cap E_2)$ is composed of the edges present in both graphs. We also define the *edge-weighted total network* of $m \leq G$ graphs $\mathcal{G}_1, \dots, \mathcal{G}_m$ as the graph formed by the union $\mathcal{G}_{1 \cup \dots \cup m} = (V, E_1 \cup \dots \cup E_m)$ where the weight $w_{jj'}$ of the edge $e_{jj'}$ is given by the cardinality of the set $\{g \in \{1, \dots, m\} : e_{jj'} \in E_g\}$. More simply, $\mathcal{G}_{1 \cup \dots \cup m}$ is determined by summing the adjacency matrices of \mathcal{G}_1 to \mathcal{G}_m . Analogously, the *signed edge-weighted total network* takes into account the stability of the sign of an edge over the classes by summing signed adjacency matrices. Naturally, the classes can also be compared by one or more summary statistics at node-, edge-, and network-level per class [cf. 29].

We also propose the idea of ‘network rewiring’. Suppose an investigator is interested in the specific interaction between genes A and B for classes g_1 and g_2 . The desire is to characterize the dependency between genes A and B and determine the differences between the two classes. To do so, we suggest using the decomposition of the covariance of A and B into the individual contributions of all paths between A and B . A path z between A and B of length t_z in a graph for class g is, following Lauritzen [24], defined to be a sequence $A = v_0, \dots, v_{t_z} = B$ of distinct vertices such that $(v_{d-1}, v_d) \in E_g$ for all $d = 1, \dots, t_z$. The possibility of the mentioned decomposition was shown by Jones and West [22] and, in terms of $\hat{\Omega}_g^0 = [\omega_{jj'}]$, can be stated as:

$$(15) \quad \text{Cov}(A, B) = \sum_{z \in \mathcal{Z}_{AB}} (-1)^{t_z+1} \omega_{Av_1} \omega_{v_1 v_2} \omega_{v_2 v_3} \cdots \omega_{v_{t_z-2} v_{t_z-1}} \omega_{v_{t_z-1} B} \frac{|(\hat{\Omega}_g^0)_{-P}|}{|\hat{\Omega}_g^0|},$$

where \mathcal{Z}_{AB} is the set of all paths between A and B and $(\hat{\Omega}_g^0)_{-P}$ denotes the matrix $\hat{\Omega}_g^0$ with rows and columns corresponding to the vertices of the path z removed. Each *term* of the covariance decomposition in (15) can be interpreted as the flow of information through a given path z between A and B in \mathcal{G}_g . Imagine performing this decomposition for A and B in both $\hat{\Omega}_{g_1}^0$ and $\hat{\Omega}_{g_2}^0$. For each path, we can then identify whether it runs through the common network $\mathcal{G}_{g_1 \cap g_2}$, or utilizes the differential networks $\mathcal{G}_{g_2 \setminus g_1}, \mathcal{G}_{g_1 \setminus g_2}$ unique to the classes. The paths that pass through the differential networks can be thought of as a ‘rewiring’ between the groups (in particular compared to the common network). In summary, the covariance between a node pair can be separated into a component that is common and a component that is differential (or rewired).

Example 4. Suppose we have the following two graphs for classes $g_1 = 1$ and $g_2 = 2$:



and consider the covariance between node A and B . In \mathcal{G}_1 the covariance $\text{Cov}(Y_A, Y_B)$ is decomposed into contributions by the paths (A, B) , $(A, 5, B)$, and $(A, 5, 4, B)$. Similarly for \mathcal{G}_2 , the contributions are from paths $(A, 5, B)$ and $(A, 5, 4, 3, B)$. Thus $(A, 5, B)$ is the only shared path. Depending on the size of the contributions we might conclude that network 1 has some ‘rewired pathways’ compared to the other. This method gives a concise overview of the estimated interactions between two given genes, which genes mediate or moderate these interactions, as well as how the interaction patterns differ across the classes. In turn this might suggest candidate genes for perturbation or knock-down experiments. ■

5. SIMULATION STUDY

In this section we explore and measure the performance of the fused estimator and its behavior in four different scenarios. Performance is measured primarily by the squared Frobenius loss,

$$L_F^{(g)}(\hat{\Omega}_g(\Lambda), \Omega_g) = \|\hat{\Omega}_g(\Lambda) - \Omega_g\|_F^2,$$

between the class precision estimate and the true population class precision matrix. However, the performance is also assessed in terms of the quadratic loss,

$$L_Q^{(g)}(\hat{\Omega}_g(\Lambda), \Omega_g) = \|\hat{\Omega}_g(\Lambda)\Omega_g^{-1} - \mathbf{I}_p\|_F^2.$$

The risk defined as the expected loss associated with an estimator, say,

$$\mathcal{R}_F\{\hat{\Omega}_g(\Lambda)\} = \mathbb{E}\left[L_F^{(g)}(\hat{\Omega}_g(\Lambda), \Omega_g)\right],$$

is robustly approximated by the median loss over a repeated number of simulations and corresponding estimations.

We designed four simulation scenarios to explore the properties and performance of the fused ridge estimator and alternatives. Scenario (1) evaluates the fused ridge estimator under two choices of the penalty matrix, the non-fused ridge estimate applied individually to the classes, and the non-fused ridge estimate using the pooled covariance matrix when (1a) $\Omega_1 = \Omega_2$ and (1b) $\Omega_1 \neq \Omega_2$. Scenario (2) evaluates the fused ridge estimator under different choices of targets: (2a) $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{0}$, (2b) $\mathbf{T}_1 = \mathbf{T}_2 = \alpha\mathbf{I}_p$, and (2c) $\mathbf{T}_1 = \mathbf{T}_2 = \Omega$. Scenario (3) evaluates the fused ridge estimator for varying network topologies and degrees of class homogeneity. Specifically, for (3a) scale-free topology and (3b) small-world topology, each with (3i) low class homogeneity and (3ii) high class homogeneity. Scenario (4) investigates the fused estimator under non-equal class sample sizes. Except for scenario 4, we make no distinction between the loss in different classes. Except for scenario 1, we use penalty matrices of the form $\Lambda = \lambda\mathbf{I}_p + \lambda_f(\mathbf{J}_p - \mathbf{I}_p)$.

5.1. Scenario 1: Fusion versus no fusion. Scenario 1 explores the loss-efficiency of the fused estimate versus non-fused estimates as a function of the class sample size n_g for fixed p and hence for different p/n_\bullet ratios. Banded population precision matrices are simulated from $G = 2$ classes. We set $p = 30$ and

$$(16) \quad (\Omega_g)_{jj'} = \frac{k+1}{|j-j'|+1} \mathbb{1}[|j-j'| \leq k]$$

with k non-zero off-diagonal bands. The sub-scenario (1a) $\Omega_1 = \Omega_2$ uses $k = 15$ bands whereas (1b) $\Omega_1 \neq \Omega_2$ uses $k = 15$ bands for Ω_1 and $k = 2$ bands for Ω_2 . Hence, identical and very different population precision matrices are considered, respectively.

For $n_g = 10, 25, 70$ the loss over 20 repeated runs was computed. In each run, the optimal *unrestricted* penalty matrix Λ was determined by LOOCV. The losses were computed for (1i) the fused ridge estimator with an unrestricted penalty matrix, (1ii) the fused ridge estimator with a restricted penalty matrix such that $\lambda_{11} = \lambda_{22}$, (1iii) the regular non-fused ridge estimator applied separately to each class, and (1iv) the regular non-fused ridge estimator using the pooled estimate \mathbf{S}_\bullet . In all cases the targets $\mathbf{T}_1 = \mathbf{T}_2 = \alpha_\bullet\mathbf{I}_p$

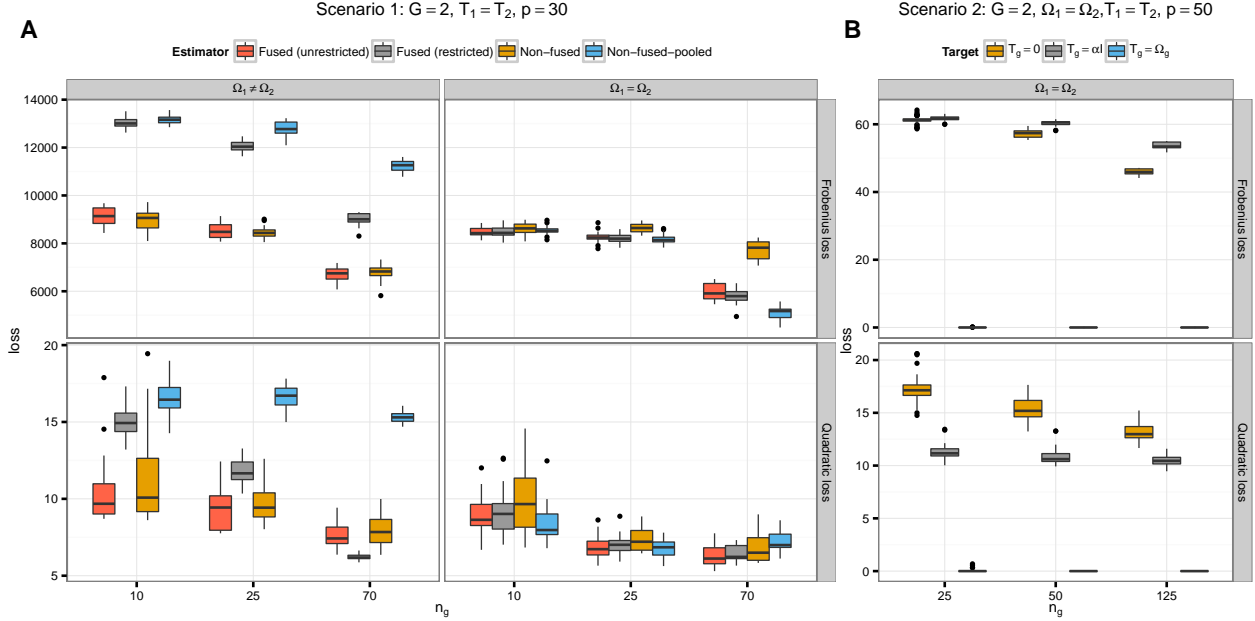


FIGURE 1. A: Results for Scenario 1. The losses against the class samples size for different ridge estimators under unequal and equal class population matrices. B: Results for Scenario 2. The losses against the class sample size with different target matrices.

were used with $\alpha_\bullet = p / \text{tr}(\mathbf{S}_\bullet)$. The risk and quartile losses for scenario 1 are seen in the boxplots of Figure 1A.

Generally, the *unrestricted* fused estimates are found to perform at least as well as the (superior of the) *non-fused* estimates. This can be expected as the fused ridge estimate might be regarded as an interpolation between using the non-fused ridge estimator on the pooled data and within each class separately. Hence, the LOOCV procedure is thus able to capture and select the appropriate penalties both when the underlying population matrices are very similar and when they are very dissimilar. In the case of differing class population precision matrices, the *restricted* fused ridge estimator (that uses the single ridge penalty $\lambda_{11} = \lambda_{22}$) performs somewhat intermediately, indicating again the added value of the flexible penalty setup. It is unsurprising that the non-fused estimate using the pooled covariance matrix is superior in scenario 1b, where $\Omega_1 = \Omega_2$, as it is the explicit estimator in this scenario, cf. Section 2.2 of the Supplementary Material.

5.2. Scenario 2: Target versus no target. Scenario 2 investigates the added value of the targeted approach to fused precision matrix estimation compared to that of setting $\mathbf{T}_g = \mathbf{0}$ which reduces to the special-case considered by Price et al. [33]. We simulated data sets with $G = 2$ classes from banded precision matrices (as given in (16)) with $p = 50$ variables and $k = 25$ bands for varying class sample sizes n_g and target matrices \mathbf{T}_1 and \mathbf{T}_2 . The performance was evaluated using (2a) $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{0}$, (2b) $\mathbf{T}_g = \alpha_\bullet \mathbf{I}_p$, as above, and (2c) the spot-on target $\mathbf{T}_1 = \mathbf{T}_2 = \Omega$ for each of $n_g = 25, 50, 125$ class sample sizes.

As above, risks were estimated by the losses for each class over 20 simulation repetitions. The optimal penalties were determined by LOOCV with penalty matrices of the form $\mathbf{A} = \lambda \mathbf{I}_p + \lambda_f (\mathbf{J}_p - \mathbf{I}_p)$. The results are shown in the boxplots in Panel B of Figure 1. As expected, the spot-on target shows superior performance in terms of loss in all cases. This suggests that well-informed choices of the target can greatly improve the estimation and that the algorithm will put emphasis on the target if it reflects the truth. Such behavior is also seen analytically in the ridge estimator of Schäfer and Strimmer [39] inferred from their closed expression of the optimal penalty. We see that using the scalar target $\alpha_\bullet \mathbf{I}_p$ results in an as-good or lower risk in terms of the quadratic but not the Frobenius loss compared to the no-target situation.

As this scenario corresponds to the case of Price et al. [33] we performed a secondary timing benchmark of their accompanying `RidgeFusion` package compared to `rag2ridges`. We evaluated estimation time of each package on a single simulated data set with $p = 50$, $G = 2$, and $n_1 = n_2 = 10$ using a banded matrix as before. The average estimation times over 100 model fits were 8.9 and 26 milliseconds for packages `rag2ridges` and `RidgeFusion`, respectively. This approximates a factor 2.94 speed-up for a single model fit. The timing was done using the package `microbenchmark` [28] and the estimates from each package were in agreement within expected numerical precision.

5.3. Scenario 3: Varying topology and class (dis)similarity. Scenario 3 investigates the fused estimator with $G = 3$ classes for (3i) high and (3ii) low class homogeneity and two different latent random graph topologies on $p = 50$ variables. The topologies are the (3a) ‘small-world’ and the (3b) ‘scale-free’ topology generated by Watts-Strogatz and Barabási graph games, respectively. The former generates topologies where all node degrees are similar while the latter game generates networks with (few) highly connected hubs. From the generated topology, we construct a latent precision matrix Ψ with diagonal elements set to 1 and the non-zero off-diagonal entries dictated by the network topology set to 0.1.

The two topologies are motivated as they imitate many real phenomena and processes. Small-world topologies approximate systems such as power grids, the neural network of the worm *C. elegans*, and the social networks of film actors [27, 44]. Conversely, scale-free topologies approximate many social networks, protein-protein interaction networks, airline networks, the world wide web, and the internet [4, 3].

We control the inter-class homogeneity using a latent inverse Wishart distribution for each class covariance matrix as considered by Bilgrau et al. [9]. That is, we let

$$(17) \quad \Sigma_g = \Omega_g^{-1} \sim \mathcal{W}_p^{-1}\left((\nu - p - 1)\Phi^{-1}, \nu\right), \quad \nu > p + 1$$

where $\mathcal{W}_p^{-1}(\Theta, \nu)$ denotes an inverse Wishart distribution with scale matrix Θ and ν degrees of freedom. The parametrization implies the expected value $\mathbb{E}[\Sigma_g] = \mathbb{E}[\Omega_g^{-1}] = \Phi^{-1}$ and thus Φ defines the latent expected topology. We simulate from a multivariate normal distribution as before conditional on the realized covariance Σ_g . In (17), the parameter ν controls the inter-class homogeneity. Large ν imply that $\Omega_1 \approx \Omega_2 \approx \Omega_3$ and thus a large class homogeneity. Small values of $\nu \rightarrow (p + 1)^+$ imply large heterogeneity.

For the simulations, we chose (i) $\nu = 100$ and (ii) $\nu = 1000$. Again we fitted the model using both the zero target as well as the scalar matrix target described above using the reciprocal value of the mean eigenvalue, i.e., $\mathbf{T}_1 = \mathbf{T}_2 = \mathbf{T}_3 = \alpha \mathbf{I}_p$ for both $\alpha = 0$ and $\alpha = p / \text{tr}(\mathbf{S}_\bullet)$. The estimation was repeated 20 times for each combination of high/low class similarity, network topology, choice of target, and class sample-size $n_1 = n_2 = n_3 = 25, 50, 125$. Panels A and B of Figure 2 show box-plots of the results.

First, the loss is seen to be dependent on the network topology, irrespective of the loss function. Second, as expected, the loss is strongly influenced by the degree of class (dis)similarity where a higher homogeneity yields a lower loss. Intuitively, this makes sense as the estimator can borrow strength across the classes and effectively increase the degrees of freedom in each class. Third, the targeted approach has a superior loss in all cases with a high class homogeneity and thus the gain in loss-efficiency is greater for the targeted approach. For low class homogeneity, the targeted approach performs comparatively to the zero target with respect to the Frobenius loss while it is seemingly better in terms of quadratic loss. Measured by quadratic loss, the targeted approach nearly always outperforms the zero target.

5.4. Scenario 4: Unequal class sizes. Scenario 4 explores the fused estimator under unequal class sample sizes. We simulated data with $k = 8$ non-zero off-diagonal bands, $G = 2$, and $p = 50$. The number of samples in class 2 was fixed at $n_2 = 30$ while the number of samples in class 1 were varied: $n_1 = 5, 25, 50, 75$. The results of the simulation are shown in Panel C of Figure 2. Note that we consider the Frobenius and quadratic loss within each class separately here.

Not surprisingly, the fused estimator performs better (for both classes) when n_\bullet increases. Perhaps more surprising, there seems to be no substantial difference in quadratic loss for group n_1 and n_2 suggesting that the fusion indeed borrows strength from the larger class. A loss difference is only visible in the most extreme case where $n_1 = 5$ and $n_2 = 30$. The relative difference however is not considered large.

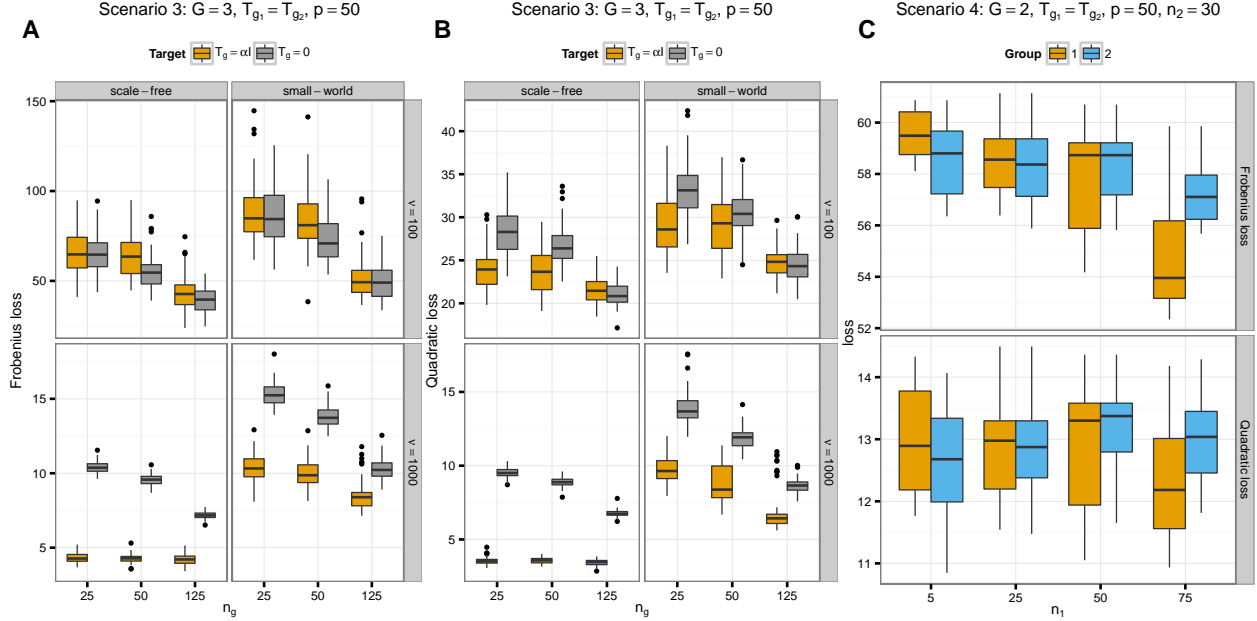


FIGURE 2. A: Scenario 3. The Frobenius loss as a function of sample size under different topologies and degree of similarity. B: Scenario 3. The quadratic loss as a function of sample size under different topologies and degree of similarity. C: Scenario 4. The loss as a function of sample size of class 1 with fixed sample size for class 2.

6. APPLICATIONS

Lymphoma refers to a group of cancers that originate in specific cells of the immune system such as white blood T- or B-cells. Approximately 90% of all lymphoma cases are non-Hodgkin's lymphomas—a diverse group of blood cancers excluding Hodgkin's disease—of which the aggressive diffuse large B-cell lymphomas (DLBCL) constitutes the largest subgroup [41]. We showcase the usage of the fused ridge estimator through two analyzes of DLBCL data.

In DLBCL, there exists at least two major genetic subtypes of tumors named after their similarities in genetic expression with activated B-cells (ABC) and germinal centre B-cells (GCB). A third *umbrella* class, usually designated as Type III, contains tumors that cannot be classified as being either of the ABC or GCB subtype. Patients with tumors of GCB class show a favorable clinical prognosis compared to that of ABC. Even though the genetic subtypes have been known for more than a decade [1] and despite the appearance of refinements to the DLBCL classification system [14], DLBCL is still treated as a singular disease in daily clinical practice and the first differentiated treatment regimens have only recently started to appear in clinical trials [36, 30]. Many known phenotypic differences between ABC and GCB are associative, which might underline the translational inertia. Hence, the biological underpinnings and *functional differences* between ABC and GCB are of central interest and the motivation for the analyzes below.

Incorrect regulation of the NF- κ B signaling pathway, responsible for i.a. control of cell survival, has been linked to cancer. This pathway has certain known drivers of deregulation. Aberrant interferon β production due to recurrent oncogenic mutations in the central MYD88 gene interferes with cell cycle arrest and apoptosis [47]. It also well-known that BCL2, another member of the NF- κ B pathway, is deregulated in DLBCL [40]. Moreover, a deregulated NF- κ B pathway is a key hallmark distinguishing the poor prognostic ABC subclass from the good prognostic GCB subclass of DLBCL [35]. Our illustrative analyzes thus focus on the *functional differences* between ABC and GCB in relation to the NF- κ B pathway. Section 6.1 investigates the DLBCL classes in the context of a single data set on the NF- κ B signalling pathway. Section 6.2 analyzes

multiple DLBCL NF- κ B data sets with a focus on finding common motifs and motif differences in network representations of pathway-deregulation. These analyzes show the value of a fusion approach to integration. In all analyzes we take the NF- κ B pathway and its constituent genes to be defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [23].

6.1. Nonintegrative analysis of DLBCL subclasses. We first analyze the data from Dybkær et al. [14], consisting of 89 DLBCL tumor samples. These samples were RMA-normalized using custom brainarray chip definition files (CDF) [12] and the R-package `affy` [20]. This preprocessing used Entrez gene identifiers (EID) by the National Center for Biotechnology Information (NCBI), which are also used by KEGG. The usage of custom CDFs avoids the mapping problems between Affymetrix probeset IDs and KEGG. Moreover, the custom CDFs can increase the robustness and precision of the expression estimates [26, 37]. The RMA-preprocessing yielded 19,764 EIDs. Subsequently, the features were reduced to the available 82 out of the 91 EIDs present in the KEGG NF- κ B pathway. The samples were then partitioned, using the DLBCL automatic classifier (DAC) by Care et al. [11], into the three classes ABC ($n_1 = 31$), III ($n_2 = 13$), and GCB ($n_3 = 45$), and gene-wise centered to have zero mean within each class.

The analysis was performed with the following settings. Target matrices for the groups were chosen to be scalar matrices with the scalar determined by the inverse of the average eigenvalue of the corresponding sample class covariance matrix, i.e.:

$$\mathbf{T}_{ABC} = \alpha_1 \mathbf{I}_p, \quad \mathbf{T}_{III} = \alpha_2 \mathbf{I}_p, \quad \mathbf{T}_{GCB} = \alpha_3 \mathbf{I}_p, \quad \text{where} \quad \alpha_g = \frac{p}{\text{tr}(\mathbf{S}_g)}.$$

These targets translate to a class-scaled ‘prior’ of conditional independence for all genes in NF- κ B. The optimal penalties were determined by LOOCV using the penalty matrix and graph given in (18). Note that the penalty setup bears resemblance to Example 2. Differing class-specific ridge penalties were allowed because of considerable differences in class sample size. Direct shrinkage between ABC and GCB was disabled by fixing the corresponding pair-fusion penalty to zero. The remaining fusion penalties were free to be estimated. Usage of the Nelder-Mead optimization procedure then resulted in the optimal values given on the right-hand side of (18) below:

$$(18) \quad \begin{array}{ccc} \text{ABC} & \text{Type III} & \text{GCB} \\ \text{---} \lambda_{11} \text{---} & \text{---} \lambda_{22} \text{---} & \text{---} \lambda_{33} \text{---} \\ & \lambda_{12} & \lambda_{23} \end{array} \quad \mathbf{\Lambda}^* = \begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 \\ \lambda_{12} & \lambda_{22} & \lambda_{23} \\ 0 & \lambda_{23} & \lambda_{33} \end{bmatrix} = \begin{bmatrix} 2 & 1.5 \times 10^{-3} & 0 \\ 1.5 \times 10^{-3} & 2.7 & 2 \times 10^{-3} \\ 0 & 2 \times 10^{-3} & 2.3 \end{bmatrix} \quad \begin{array}{l} \text{ABC} \\ \text{III} \\ \text{GCB} \end{array}$$

The ridge penalties of classes ABC and GCB are seen to be comparable in size. The small size of the Type III class leads to a relatively larger penalty to ensure a well-conditioned and stable estimate. The estimated fusion penalties are all relatively small, implying that heavy fusion is undesirable due to class-differences. The three class-specific precision matrices were estimated under $\mathbf{\Lambda}^*$ and subsequently scaled to partial correlation matrices. Panels A–C of Figure 3 visualize these partial correlation matrices. In general, the ABC and GCB classes seem to carry more signal in both the negative and positive range vis-à-vis the Type III class.

Post-hoc support determination was carried out on the partial correlation matrices using the class-wise lFDR approach of Section 4.3. The lFDR threshold was chosen conservatively to 0.99, selecting 39, 85, 34 edges for classes ABC, III, GCB, respectively. The relatively high number of edges selected for the Type III class is (at least partly) due to the difficulty of determining the mixture distribution mentioned in Section 4.3 when the overall partial correlation signal is relatively flat. Panels D–E of Figure 3 then show the conditional independence graphs corresponding to the sparsified partial correlation matrices. We note that a single connected component is identified in each class, suggesting, at least for the ABC and GCB classes, a genuine biological signal. A secondary supporting overview is provided in Table 1.

Table 1 gives the most central genes in the graphs of Panels D–E by two measures of node centrality: degree and betweenness. The node degree indicates the number of edges incident upon a particular node. The betweenness centrality indicates in how many shortest paths between vertex pairs a particular node acts as an intermediate vertex. Both measures are proxies for the importance of a feature. See, e.g., [29] for an overview of these and other centrality measures. It is seen that the CCL, CXCL, and TNF gene families

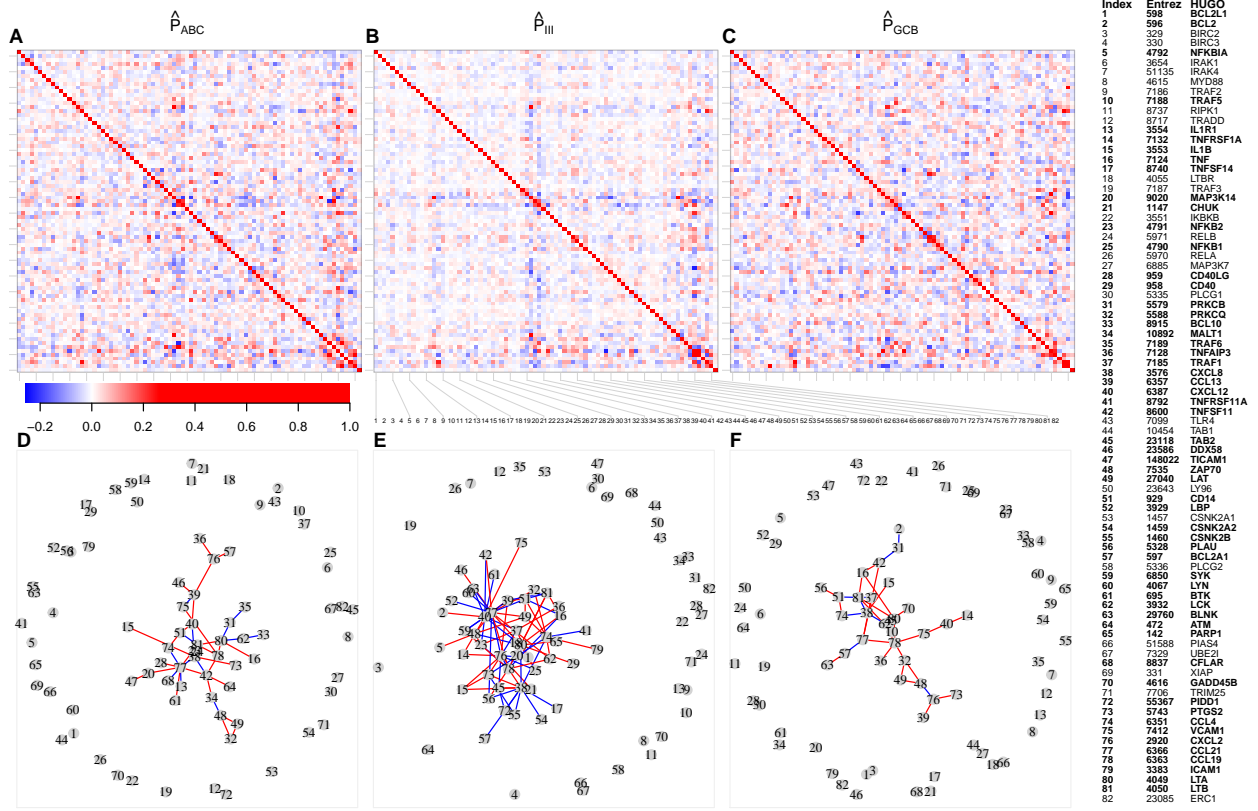


FIGURE 3. *Top*: Heat maps and color key of the partial correlation matrices for the ABC (panel A), III (panel B), and GCB (panel C) classes in the NF- κ B signaling pathway on the Dybkær et al. [14] data. *Bottom*: Graphs corresponding to the sparsified precision matrices for the classes above. Red and blue edges correspond to positive and negative partial correlations, respectively. *Far right-panel*: EID key and corresponding Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) curated gene names of the NF- κ B signaling pathway genes. Genes that are connected in panels D–F are shown bold.

are well-represented as central and connected nodes across all classes. The gene CCL21 is very central in classes ABC and III, but less so in the GCB class. From Panels D–E of Figure 3 it is seen that BCL2 and BCL2A1 are only connected in the non-ABC classes. Contrary to expectation, MYD88 is disconnected in all graphs. The genes ZAP70, LAT, and LCK found in Figure 3 and Table 1 are well-known T-cell specific genes involved in the initial T-cell receptor-mediated activation of NF- κ B in T-cells [7]. From the differences in connectivity of these genes, different abundances of activated T-cells or different NF- κ B activation programs for ABC/GCB might be hypothesized.

6.2. Integrative DLBCL analysis. We now expand the analysis of the previous section to show the advantages of integration by fusion. A large number of DLBCL gene expression profile (GEP) data sets is freely available at the NCBI Gene Expression Omnibus (GEO) website [5]. We obtained 11 large-scale DLBCL data sets whose GEO-accession numbers (based on various Affymetrix microarray platforms) can be found in the first column of Table 2. One of the sets, with GEO-accession number GSE11318, is treated as a pilot/training data set for the construction of target matrices (see below). The GSE10846 set is composed of two distinct data sets corresponding to two treatment regimens (R-CHOP and CHOP) as well as different

TABLE 1. The most central genes, their EID, and their plot index. For each class and node, the degree (with the number of positive and negative edges connected to that node in parentheses) and the betweenness centrality is shown. Only the 15 genes with the highest degrees summed over each class are shown.

	EID	Index	ABC		III		GCB	
			Degree	Betw.	Degree	Betw.	Degree	Betw.
CCL21	6366	77	9 (5 ⁺ , 4 ⁻)	202.0	17 (9 ⁺ , 8 ⁻)	297.00	4 (3 ⁺ , 1 ⁻)	106
CXCL8	3576	38	5 (2 ⁺ , 3 ⁻)	126.0	12 (4 ⁺ , 8 ⁻)	234.00	4 (1 ⁺ , 3 ⁻)	56
CCL19	6363	78	4 (4 ⁺ , 0 ⁻)	120.0	10 (6 ⁺ , 4 ⁻)	91.70	6 (6 ⁺ , 0 ⁻)	230
LTA	4049	80	5 (3 ⁺ , 2 ⁻)	143.0	10 (6 ⁺ , 4 ⁻)	195.00	3 (3 ⁺ , 0 ⁻)	56
CXCL12	6387	40	3 (2 ⁺ , 1 ⁻)	84.2	12 (5 ⁺ , 7 ⁻)	187.00	2 (2 ⁺ , 0 ⁻)	27
CXCL2	2920	76	3 (3 ⁺ , 0 ⁻)	61.0	11 (5 ⁺ , 6 ⁻)	196.00	3 (2 ⁺ , 1 ⁻)	53
LTB	4050	81	4 (3 ⁺ , 1 ⁻)	85.5	5 (3 ⁺ , 2 ⁻)	4.24	6 (3 ⁺ , 3 ⁻)	98
CD14	929	51	3 (2 ⁺ , 1 ⁻)	20.2	6 (3 ⁺ , 3 ⁻)	25.90	3 (2 ⁺ , 1 ⁻)	32
CCL4	6351	74	2 (1 ⁺ , 1 ⁻)	5.0	8 (5 ⁺ , 3 ⁻)	118.00	2 (1 ⁺ , 1 ⁻)	4
ZAP70	7535	48	3 (2 ⁺ , 1 ⁻)	60.0	5 (4 ⁺ , 1 ⁻)	50.70	3 (2 ⁺ , 1 ⁻)	75
CCL13	6357	39	4 (3 ⁺ , 1 ⁻)	119.0	5 (3 ⁺ , 2 ⁻)	19.70	1 (1 ⁺ , 0 ⁻)	0
TNFSF11	8600	42	5 (4 ⁺ , 1 ⁻)	160.0	2 (1 ⁺ , 1 ⁻)	0.00	3 (2 ⁺ , 1 ⁻)	55
TNF	7124	16	1 (1 ⁺ , 0 ⁻)	0.0	4 (2 ⁺ , 2 ⁻)	1.68	3 (3 ⁺ , 0 ⁻)	24
LAT	27040	49	2 (2 ⁺ , 0 ⁻)	0.0	4 (4 ⁺ , 0 ⁻)	15.80	2 (2 ⁺ , 0 ⁻)	0
LCK	3932	62	2 (0 ⁺ , 2 ⁻)	31.0	3 (3 ⁺ , 0 ⁻)	10.00	3 (2 ⁺ , 1 ⁻)	64

TABLE 2. Overview of data sets, the defined classes, and the number of samples. In GSE31312, 28 samples were not classified with the DAC due to technical issues and hence do not appear in this table. In the pilot study GSE11318, 31 samples were primary mediastinal B-cell lymphoma and left out. Note also that the pilot data set GSE11318 was not classified by the DAC.

	ABC		Type III		GBC		$\sum n_g$
	g	n_g	g	n_g	g	n_g	
Pilot data							
GSE11318		74		71		27	172
Data set							
GSE56315	1	31	2	13	3	45	89
GSE19246	4	51	5	30	6	96	177
GSE12195	7	40	8	18	9	78	136
GSE22895	10	31	11	21	12	49	101
GSE31312	13	146	14	97	15	224	467
GSE10846.CHOP	16	64	17	28	18	89	181
GSE10846.RCHOP	19	75	20	42	21	116	233
GSE34171.hgu133plus2	22	23	23	15	24	52	90
GSE34171.hgu133AplusB	25	18	26	17	27	43	78
GSE22470	28	86	29	43	30	142	271
GSE4475	31	73	32	20	33	128	221
$\sum n_g$		638		344		1062	2044

time-periods of study. Likewise, GSE34171 is composed of three data sets corresponding to the respective microarray platforms used: HG-U133A, HG-U133B, and HG-U133 plus 2.0. As the samples on HG-U133A and HG-U133B were paired and run on *both* platforms, the (overlapping) features were averaged to form a single virtual microarray comparable to that of HG-U133 plus 2.0. Note that the Dybkær et al. [14] data used in Section 6.1 is part of the total batch under GEO-accession number GSE56315. The sample sizes for

the individual data sets vary in the range 78–495 and can also be found in Table 2. The data yield a total of 2,276 samples making this, to our knowledge, the hitherto largest integrative DLBCL study.

Similar to above, all data sets were RMA-normalized using custom brainarray CDFs and the R-package **affy**. Again, NCBI EIDs were used to avoid non-bijective gene-ID translations between the array-platforms and the KEGG database. The freely available R-package **DLBCLdata** was created to automate the download and preprocessing of the data sets in a reproducible and convenient manner. See the **DLBCLdata** documentation [8] for more information. Subsequently, the data sets were reduced to the intersecting 11,908 EIDs present on all platforms. All samples in all data sets, except for the pilot study GSE11318, were classified as either ABC, GCB, or Type III using the DAC mentioned above. The same classifier was used in all data sets to obtain a uniform classification scheme and thus maximize the comparability of the classes across data sets. Subsequently, the features were reduced to the EIDs present in the NF- κ B pathway and gene-wise centered to have zero mean within each combination of DLBCL subtype and data set. We thus have a two-way study design—DLBCL subtypes and multiple data sets—analogue to Example 3. A concise overview of each of the $11 \times 3 = 33$ classes for the non-pilot data is provided in Table 2.

The target matrices were constructed from the pilot data in an attempt to use information in the directed representation \mathcal{G}_{pw} of the NF- κ B pathway obtained from KEGG. The directed graph represents direct and indirect causal interactions between the constituent genes. It was obtained from the KEGG database via the R-package **KEGGgraph** [50]. A target matrix was constructed for each DLBCL subtype using the pilot data and the information from the directed topology by computing node contributions using multiple linear regression models. That is, from an initial $\mathbf{T} = \mathbf{0}$, we update \mathbf{T} for each node $\alpha \in V(\mathcal{G}_{pw})$ through the following sequence:

$$\begin{aligned} T_{\alpha,\alpha} &:= T_{\alpha,\alpha} + \frac{1}{\sigma^2} \\ \mathbf{T}_{pa(\alpha),\alpha} &:= \mathbf{T}_{pa(\alpha),\alpha} + \frac{1}{\sigma^2} \boldsymbol{\beta}_{pa(\alpha)} \\ \mathbf{T}_{\alpha,pa(\alpha)} &:= \mathbf{T}_{\alpha,pa(\alpha)} + \frac{1}{\sigma^2} \boldsymbol{\beta}_{pa(\alpha)} \\ \mathbf{T}_{pa(\alpha),pa(\alpha)} &:= \mathbf{T}_{pa(\alpha),pa(\alpha)} + \frac{1}{\sigma^2} \boldsymbol{\beta}_{pa(\alpha)} \boldsymbol{\beta}_{pa(\alpha)}^\top, \end{aligned}$$

where $pa(\alpha)$ denotes the parents of node α in \mathcal{G}_{pw} , and where σ and $\boldsymbol{\beta}$ are the residual standard error and regression coefficients obtained from the linear regression of α on $pa(\alpha)$. By this scheme the target matrix represents the conditional independence structure that would result from moralizing the directed graph. If \mathcal{G}_{pw} is acyclic then $\mathbf{T} \succ 0$ is guaranteed.

The penalty setup bears resemblance to Example 3. The Type III class is considered closer to the ABC and GCB subtypes than ABC is to GCB. Thus, the direct shrinkage between the ABC and GCB subtypes was fixed to zero. Likewise, direct shrinkage between subtype and data set combinations was also disabled. Hence, a common ridge penalty λ , a data set–data set shrinkage parameter λ_{DS} and a subtype–subtype shrinkage parameter λ_{ST} were estimated. The optimal penalties were determined by SLOOCV using the penalty matrix and graph given in (19) below:

(19)

$$\mathbf{A} = \begin{bmatrix} \lambda & \lambda_{ST} & 0 & \lambda_{DS} & 0 & 0 & \dots & \lambda_{DS} & 0 & 0 \\ \lambda_{ST} & \lambda & \lambda_{ST} & 0 & \lambda_{DS} & 0 & \dots & 0 & \lambda_{DS} & 0 \\ 0 & \lambda_{ST} & \lambda & 0 & 0 & \lambda_{DS} & \dots & 0 & 0 & \lambda_{DS} \\ \lambda_{DS} & 0 & 0 & \lambda & \lambda_{ST} & 0 & \dots & \lambda_{DS} & 0 & 0 \\ 0 & \lambda_{DS} & 0 & \lambda_{ST} & \lambda & \lambda_{ST} & \dots & 0 & \lambda_{DS} & 0 \\ 0 & 0 & \lambda_{DS} & 0 & \lambda_{ST} & \lambda & \dots & 0 & 0 & \lambda_{DS} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \lambda_{DS} & 0 & 0 & \lambda_{DS} & 0 & 0 & \dots & \lambda & \lambda_{ST} & 0 \\ 0 & \lambda_{DS} & 0 & 0 & \lambda_{DS} & 0 & \dots & \lambda_{ST} & \lambda & \lambda_{ST} \\ 0 & 0 & \lambda_{DS} & 0 & 0 & \lambda_{DS} & \dots & 0 & \lambda_{ST} & \lambda \end{bmatrix}$$

The optimal penalties were found to be $\lambda^\diamond = 2.2$ for the ridge penalty, $\lambda_{DS}^\diamond = 0.0022$ for the data set fusion penalty, and $\lambda_{ST}^\diamond = 6.8 \times 10^{-4}$ for the subtype fusion penalty, respectively.

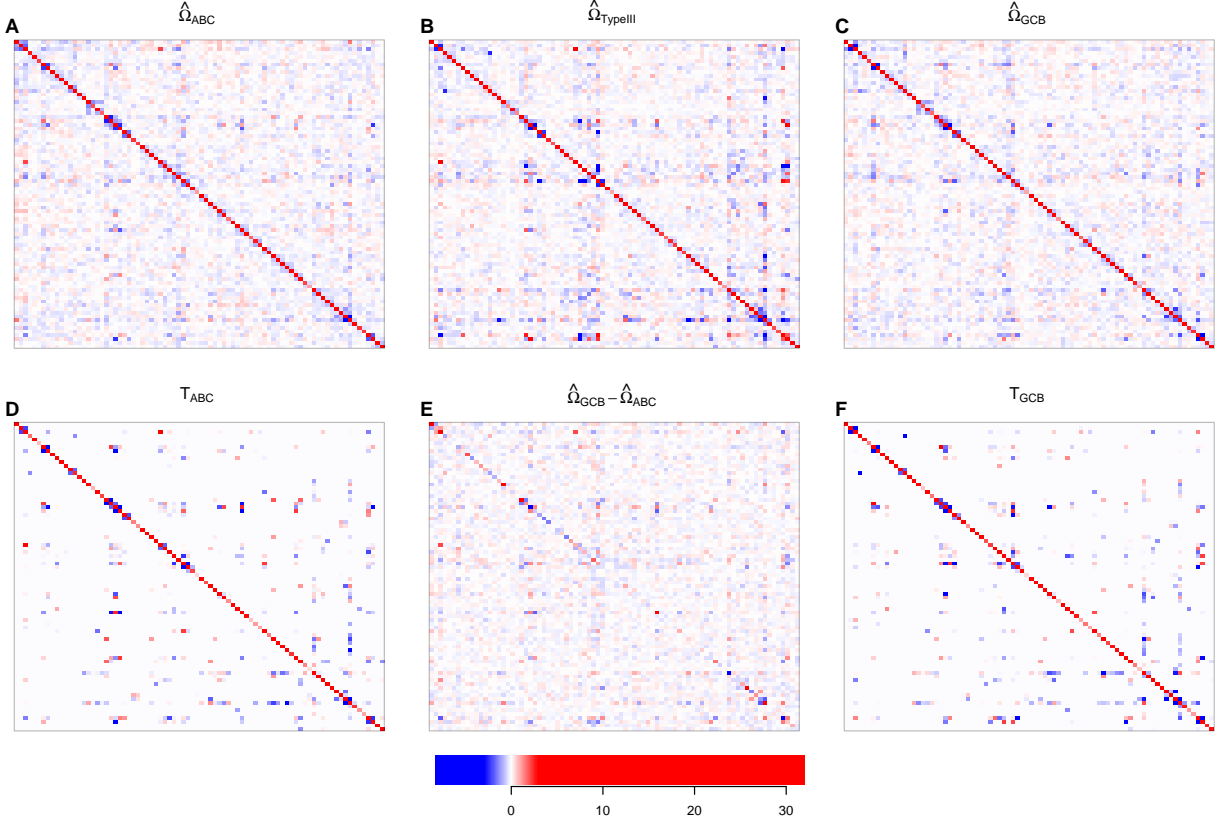


FIGURE 4. Summary of the estimated precision matrices for the NF- κ B pathway. *Top row:* Heat maps of the estimated precision matrices pooled across data sets for each genetic subtype. *Middle row from left to right:* The pooled target matrix for ABC, the difference between the pooled ABC and GCB estimates, and the pooled target matrix for GCB. *Bottom:* The color key for the heat maps.

To summarize and visualize the 33 class precision estimates they were pooled within DLBCL subtype. Panels A–C of Figure 4 visualizes the 3 pooled estimates as heat maps. Panels D and F visualize the constructed target matrices for the ABC and GCB subtypes, respectively. Panel E then gives the difference between the pooled ABC and GCB estimates, indicating that they harbor differential signals to some degree. We would like to capture the commonalities and differences with a differential network representation.

The estimated class-specific precision matrices were subsequently scaled to partial correlation matrices. Each precision matrix was then sparsified using the lFDR procedure of Section 4.3. Given the class an edge was selected whenever $1 - \widehat{\text{lFDR}} \geq 0.999$. To compactly visualize the the multiple GGMs we obtained *signed edge-weighted total networks* mentioned in Section 4.4. Clearly, for inconsistent connections the weight would vary around zero, while edges that are consistently selected as positive (negative) will have a large positive (negative) weight. These meta-graphs are plotted in Figure 5. Panels A–C give the signed edge-weighted total networks for each subtype across the data sets. They show that (within DLBCL subtypes) there are a number of edges that are highly concordant across all data sets. To evaluate the greatest differences between the ABC and GCB subtypes, the signed edge-weighted total network of the latter was subtracted from the former. The resulting graph $\mathcal{G}_{\text{ABC-GCB}}$ can be found in Panel D. Edges that are more stably present in the ABC subtype are represented in orange and the edges more stably present in the GCB subtype are represented in blue. Panel F represents the graph from panel D with only those edges retained whose

absolute weight exceeds 2. In a sense, the graph of panel F then represents the stable differential network. The strongest connections here should suggest places of regulatory deregulation gained or lost across the two subtypes. Interestingly, this differential network summary shows relatively large connected subgraphs suggesting differing regulatory mechanisms.

The graph in panel F of Figure 5 then conveys the added value of the integrative fusion approach. Certain members of the CCL, CXCL, and TNF gene families who were highly central in the analysis of Section 6.1 are still considered to be central here. However, it is also seen that certain genes that garnered high centrality measures in the single data set analyzed in Section 6.1 do not behave stably *across* data sets, such as CXCL2. In addition, the integrative analysis appoints the BCL2 gene family a central role, especially in relation to the ABC subtype. This contrasts with Section 6.1, where the BCL2 gene family was not considered central and appeared to be connected mostly in the non-ABC classes. Moreover, whereas the analysis of the single data set could not identify a signal for MYD88, the integrative analysis identifies MYD88 to be stably connected across data sets. Especially the latter two observations are in line with current knowledge on deregulation in the NF- κ B pathway in DLBCL patients. Also in accordance with literature is the known interaction of LTA with LTB seen in panel F of Figure 5 [45, 10] which here appear to be differential between ABC/GCB. Thus, borrowing information across classes enables a meta-analytic approach that can uncover information otherwise unobtainable through the analysis of single data sets.

7. DISCUSSION AND CONCLUSION

We considered the problem of jointly estimating multiple inverse covariance matrices from high-dimensional data consisting of distinct classes. A fused ridge estimator was proposed that generalizes previous contributions in two principal directions. First, we introduced the use of targets in fused ridge precision estimation. The targeted approach helps to stabilize the estimation procedure and allows for the incorporation of prior knowledge. It also juxtaposes itself with various alternative penalized precision matrix estimators that pull the estimates towards the edge of the parameter space, i.e., who shrink towards the non-interpretable null matrix. Second, instead of using a single ridge penalty and a single fusion penalty parameter for all classes, the approach grants the use of *class-specific* ridge penalties and *class-pair-specific* fusion penalties. This results in a flexible shrinkage framework that (i) allows for class-specific tuning, that (ii) supports analyzes when a factorial design underlies the available classes, and that (iii) supports the appropriate handling of situations where some classes are high-dimensional whilst others are low-dimensional. Targeted shrinkage and usage of a flexible penalty matrix might also benefit other procedures for precision matrix estimation such as the fused graphical lasso [13].

The targeted fused ridge estimator was combined with post-hoc support determination, which serves as a basis for integrative or meta-analytic Gaussian graphical modeling. This combination thus has applications in meta-, integrative-, and differential network analysis of multiple data sets or classes of data. This meta-approach to network analysis has multiple motivations. First, by combining data it can effectively increase the sample size in settings where samples are relatively scarce or expensive to produce. In a sense it refocuses the otherwise declining attention to obtaining a sufficient amount of data—a tendency we perceive to be untenable. Second, aggregation across multiple data sets decreases the likelihood of capturing idiosyncratic features (of individual data sets), thereby preventing over-fitting of the data.

Insightful summarization of the results is important for the feasibility of our approach to fused graphical modeling. To this end we have proposed various basic tools to summarize commonalities and differences over multiple graphs. These tools were subsequently used in a differential network analysis of the NF- κ B signaling pathway in DLBCL subtypes over multiple GEP data sets. This application is not without critique, as it experiences a problem present in many GEP studies: The classification of the DLBCL subtypes (ABC and GCB) is performed on the basis of the same GEP data on which the network analysis is executed. This may be deemed methodologically undesirable. However, we justify this double use of data as (a) the pathway of interest involves a selection of genes whereas the classification uses all genes, and (b) the analysis investigates partial correlations and differential networks whereas the classification, in a sense, considers only differential expression. Furthermore, as in all large-scale genetic screenings, the analyzes should be considered ‘tentative’ and findings need to be validated in independent experiments. Notwithstanding, the analyzes show that

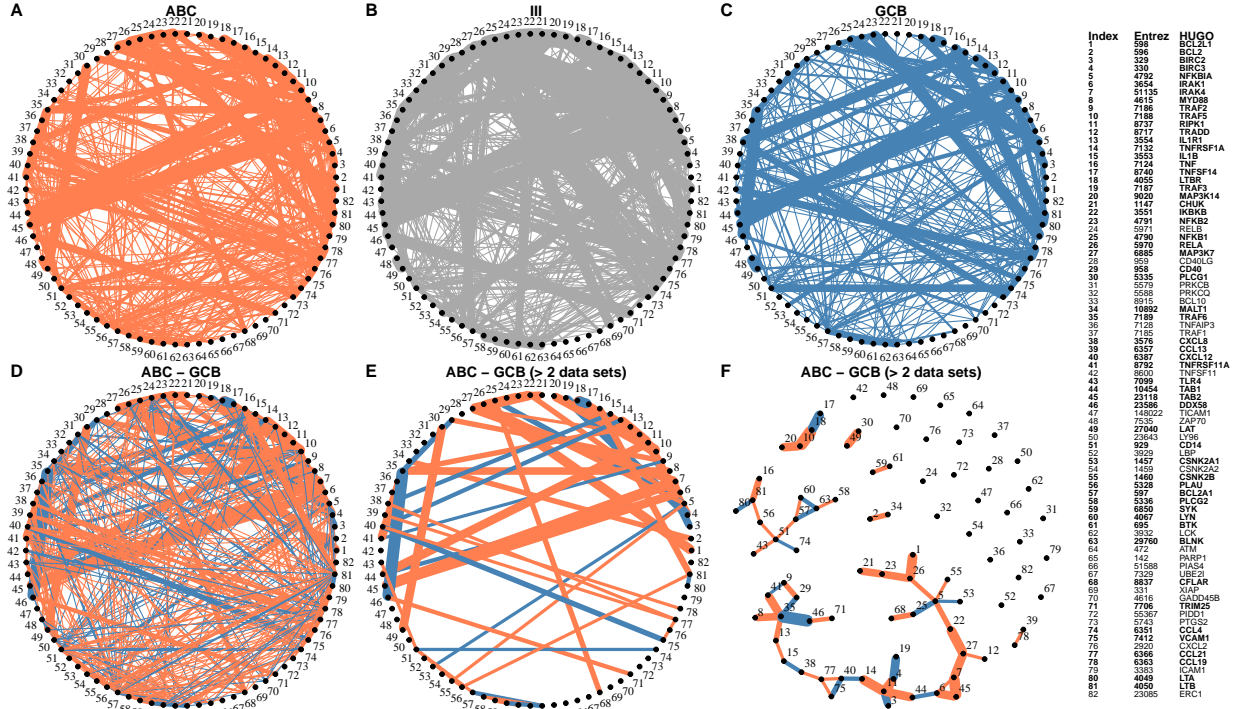


FIGURE 5. Summary of estimated GGMs for the NF- κ B pathway. *Panels A-C*: Graphs obtained by adding the signed adjacency matrices for each subtype across the data sets. The edge widths are drawn proportional to the absolute edge weight. *Panel D*: Graph obtained by subtracting the summarized signed adjacency matrix of GCB (panel A) from that of ABC (panel C). Edge widths are drawn proportional to the absolute weight and colored according to the sign. Orange implies edges more present in ABC and blue implies edges more present in GCB. *Panel E*: As the graph in panel D, however only edges with absolute weight > 2 are drawn. *Panel F*: As the graph in panel E, but with an alternative layout. *Far right-panel*: EID key and corresponding HGNC curated gene names of the NF- κ B pathway genes. Genes that are connected in panel F are shown bold.

the fusion approach to network integration has merit in uncovering class-specific information on pathway deregulation. Moreover, they exemplify the exploratory *hypothesis generating* thrust of the framework we offer.

We see various inroad for further research. With regard to estimation one could think of extending the framework to incorporate a fused version of the elastic net. Mixed fusion, in the sense that one could do graphical lasso estimation with ridge fusion or ridge estimation with lasso fusion, might also be of interest. From an applied perspective the desire is to expand the toolbox for insightful (visual) summarization of commonalities and differences over multiple graphs. Moreover, it is of interest to explore improved ways for support determination. The lFDR procedure, for example, could be expanded by considering all classes jointly. Instead of applying the lFDR procedure to each class-specific precision matrix, one would then be interested in determining the proper mixture of a grand common null-distribution and multiple class-specific non-null distributions. These inroads were out of the scope of current work, but we hope to explore them elsewhere.

7.1. Software implementation. The fused ridge estimator and its accompanying estimation procedure is implemented in the `rags2ridges`-package [31] for the statistical language R. This package has many

supporting functions for penalty parameter selection, graphical modeling, as well as network analysis. We will report on its full functionality elsewhere. The package is freely available from the Comprehensive R Archive Network: <http://cran.r-project.org/>.

ACKNOWLEDGEMENTS

Anders E. Bilgrau was supported by a grant from the Karen Elise Jensen Fonden, a travel grant from the Danish Cancer Society, and a visitor grant by the Dept. of Mathematics of the VU University Amsterdam. Carel F.W. Peeters received funding from the European Community's Seventh Framework Programme (FP7, 2007-2013), Research Infrastructures action, under grant agreement No. FP7-269553 (EpiRadBio project). The authors would also like to thank Karen Dybkær of the Dept. of Haematology at Aalborg University Hospital, for her help on the biological interpretations in the DLBCL application.

APPENDIX A. GEOMETRIC INTERPRETATION OF THE FUSED RIDGE PENALTY

Some intuition behind the fused ridge is provided by pointing to the equivalence of penalized and constrained optimization. To build this intuition we study the geometric interpretation of the fused ridge penalty in the special case of (6) with $\mathbf{T} = \mathbf{0}$. In this case $\lambda_{gg} = \lambda$ for all g , and $\lambda_{g_1 g_2} = \lambda_f$ for all $g_1 \neq g_2$. Clearly, the penalty matrix then amounts to $\mathbf{\Lambda} = \lambda \mathbf{I}_p + \lambda_f (\mathbf{J}_p - \mathbf{I}_p)$. Matters are simplified further by considering $G = 2$ classes and by focusing on a specific entry in the precision matrix, say $(\mathbf{\Omega}_g)_{jj'} = \omega_{jj'}^{(g)}$, for $g = 1, 2$. By doing so we ignore the contribution of other precision elements to the penalty. Now, the fused ridge penalty may be rewritten as:

$$\frac{\lambda}{2} (\|\mathbf{\Omega}_1\|_F^2 + \|\mathbf{\Omega}_2\|_F^2) + \frac{\lambda_f}{4} \sum_{g_1=1}^2 \sum_{g_2=1}^2 \|\mathbf{\Omega}_{g_1} - \mathbf{\Omega}_{g_2}\|_F^2 = \frac{\lambda}{2} (\|\mathbf{\Omega}_1\|_F^2 + \|\mathbf{\Omega}_2\|_F^2) + \frac{\lambda_f}{2} \|\mathbf{\Omega}_1 - \mathbf{\Omega}_2\|_F^2.$$

Subsequently considering only the contribution of the $\omega_{jj'}^{(g)}$ entries implies this expression can be further reduced to:

$$\frac{\lambda}{2} [(\omega_{jj'}^{(1)})^2 + (\omega_{jj'}^{(2)})^2] + \frac{\lambda_f}{2} (\omega_{jj'}^{(1)} - \omega_{jj'}^{(2)})^2 = \frac{\lambda + \lambda_f}{2} [(\omega_{jj'}^{(1)})^2 + (\omega_{jj'}^{(2)})^2] - \lambda_f \omega_{jj'}^{(1)} \omega_{jj'}^{(2)}.$$

It follows immediately that this penalty imposes constraints on the parameters $\omega_{jj'}^{(1)}$ and $\omega_{jj'}^{(2)}$, amounting to the set:

$$(20) \quad \left\{ (\omega_{jj'}^{(1)}, \omega_{jj'}^{(2)}) \in \mathbb{R}^2 : \frac{\lambda + \lambda_f}{2} [(\omega_{jj'}^{(1)})^2 + (\omega_{jj'}^{(2)})^2] - \lambda_f \omega_{jj'}^{(1)} \omega_{jj'}^{(2)} \leq c \right\},$$

for some $c \in \mathbb{R}_+$. It implies that the fused ridge penalty can be understood by the implied constraints on the parameters. Figure 6 shows the boundary of the set for selected values.

Panel 6A reveals the effect of the fused, inter-class penalty parameter λ_f (while keeping λ fixed). At $\lambda_f = 0$, the constraint coincides with the regular ridge penalty. As λ_f increases, the ellipsoid shrinks along the minor principal axis $x = -y$ with no shrinkage along $x = y$. In the limit $\lambda_f \rightarrow \infty$ the ellipsoid collapses onto the identity line. Hence, the parameters $\omega_{jj'}^{(1)}$ and $\omega_{jj'}^{(2)}$ are shrunken towards each other and while their differences vanish, their sum is not affected. Hence, the fused penalty parameter primarily shrinks the ‘sum of the parameters’, but also fuses them as a bound on their sizes implies a bound on their difference.

Panel 6B shows the effect of the intra-class λ penalty (while keeping λ_f fixed). When the penalty vanishes for $\lambda \rightarrow 0$ the domain becomes a degenerated ellipse (i.e. cylindrical for more than 2 classes) and parameters $\omega_{jj'}^{(1)}$ and $\omega_{jj'}^{(2)}$ may assume any value as long as their difference is less than $\sqrt{2c/\lambda_f}$. For any $\lambda > 0$, the parameter-constraint is ellipsoidal. As λ increases the ellipsoid is primarily shrunken along the principal axis formed by the identity line and along the orthogonal principal axis ($y = -x$). In the limit $\lambda \rightarrow \infty$ the ellipsoid collapses onto the point (0,0). It is clear that the shape of the domain in (20) is only determined by the ratio of λ and λ_f .

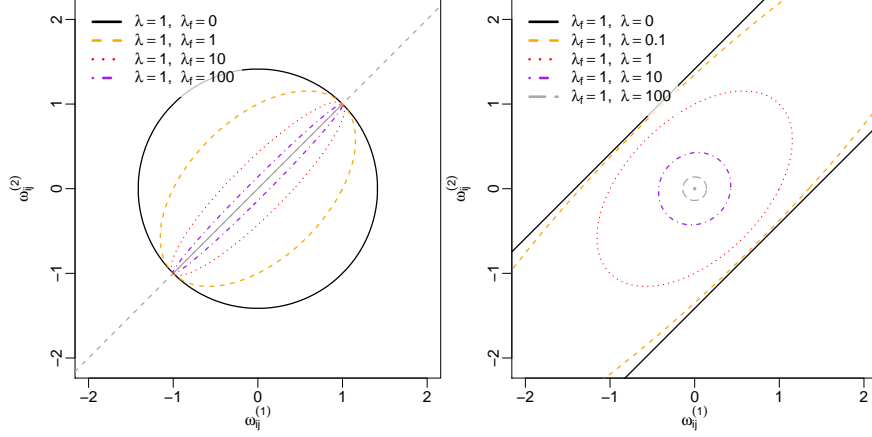


FIGURE 6. Visualization of the effects of the fused ridge penalty in terms of constraints. The left panel shows the effect of λ_f for fixed λ . Here, $\lambda_f = 0$ is the regular ridge penalty. The right panel shows the effect of λ while keeping λ_f fixed.

The effect of the penalties on the domain of the obtainable estimates can be further understood by noting that the fused ridge penalty (4) can be rewritten as

$$(21) \quad \tilde{\lambda} \sum_{g_1, g_2} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) + (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2 + \tilde{\lambda}_f \sum_{g_1, g_2} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2,$$

for some penalties $\tilde{\lambda}$ and $\tilde{\lambda}_f$. The details of this derivation can be found in Section A.1 below. The first and second summand of the rewritten penalty (21) respectively shrink the sum and difference of the parameters of the precision matrices. Their contributions thus coincide with the principal axes along which two penalty parameters shrink the domain of the parameters.

A.1. Alternative form for the fused ridge penalty. This section shows that the alternative form (21) for the ridge penalty can be written in the form (4). We again assume a common ridge penalty $\lambda_{gg} = \lambda$ and a common fusion penalty $\lambda_{g_1 g_2} = \lambda_f$ for all classes and pairs thereof. To simplify the notation, let $\mathbf{A}_g = \mathbf{\Omega}_g - \mathbf{T}_g$. Now,

$$\begin{aligned} & f^{\text{FR}'}(\{\mathbf{\Omega}_g\}; \tilde{\lambda}, \tilde{\lambda}_f, \{\mathbf{T}_g\}) \\ &= \tilde{\lambda} \sum_{g_1, g_2} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) + (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2 + \tilde{\lambda}_f \sum_{g_1, g_2} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2 \\ &= \tilde{\lambda} \sum_{g_1, g_2} \|\mathbf{A}_{g_1} + \mathbf{A}_{g_2}\|_F^2 + \tilde{\lambda}_f \sum_{g_1, g_2} \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 \\ &= \tilde{\lambda} \sum_{g_1, g_2} \left(\|\mathbf{A}_{g_1}\|_F^2 + \|\mathbf{A}_{g_2}\|_F^2 + 2\langle \mathbf{A}_{g_1}, \mathbf{A}_{g_2} \rangle \right) + \tilde{\lambda}_f \sum_{g_1, g_2} \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 \\ &= \tilde{\lambda} \sum_{g_1, g_2} \left(2\|\mathbf{A}_{g_1}\|_F^2 + 2\|\mathbf{A}_{g_2}\|_F^2 - \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 \right) + \tilde{\lambda}_f \sum_{g_1, g_2} \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 \\ &= 4\tilde{\lambda}G \sum_g \|\mathbf{A}_g\|_F^2 - \tilde{\lambda} \sum_{g_1, g_2} \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 + \tilde{\lambda}_f \sum_{g_1, g_2} \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 \\ &= 4\tilde{\lambda}G \sum_g \|\mathbf{A}_g\|_F^2 + (\tilde{\lambda}_f - \tilde{\lambda}) \sum_{g_1, g_2} \|\mathbf{A}_{g_1} - \mathbf{A}_{g_2}\|_F^2 \\ &= 4\tilde{\lambda}G \sum_g \|(\mathbf{\Omega}_g - \mathbf{T}_g)\|_F^2 + (\tilde{\lambda}_f - \tilde{\lambda}) \sum_{g_1, g_2} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2. \end{aligned}$$

Hence, the alternative penalty (21) is also of the form (4) and thus the fused ridge of (21) is equivalent to (4) for appropriate choices of the penalties.

APPENDIX B. RESULTS AND PROOFS

Section B.1 contains supporting results from other sources and results in support of Algorithm 1. Section B.2 contains proofs of the results stated in the main text as well as additional results conducive in those proofs.

B.1. Supporting results.

Lemma 1 (van Wieringen and Peeters [42]). *Amend the log-likelihood (1) with the ℓ_2 -penalty*

$$\frac{\lambda}{2} \|\mathbf{\Omega} - \mathbf{T}\|_F^2,$$

with $\mathbf{T} \in \mathcal{S}_+^p$ denoting a fixed symmetric p.s.d. target matrix, and where $\lambda \in (0, \infty)$ denotes a penalty parameter. The zero gradient equation w.r.t. the precision matrix then amounts to

$$(22) \quad \hat{\mathbf{\Omega}}^{-1} - (\mathbf{S} - \lambda \mathbf{T}) - \lambda \hat{\mathbf{\Omega}} = \mathbf{0},$$

whose solution gives a penalized ML ridge estimator of the precision matrix:

$$\hat{\mathbf{\Omega}}(\lambda) = \left\{ \left[\lambda \mathbf{I}_p + \frac{1}{4} (\mathbf{S} - \lambda \mathbf{T})^2 \right]^{1/2} + \frac{1}{2} (\mathbf{S} - \lambda \mathbf{T}) \right\}^{-1}.$$

Lemma 2 (van Wieringen and Peeters [42]). *Consider $\hat{\mathbf{\Omega}}(\lambda)$ from Lemma 1 and define $[\hat{\mathbf{\Omega}}(\lambda)]^{-1} \equiv \hat{\mathbf{\Sigma}}(\lambda)$. The following identity then holds:*

$$\mathbf{S} - \lambda \mathbf{T} = \hat{\mathbf{\Sigma}}(\lambda) - \lambda \hat{\mathbf{\Omega}}(\lambda).$$

Lemma 3. *Let $\mathbf{\Lambda} \in \mathcal{S}^G$ be a matrix of fixed penalty parameters such that $\mathbf{\Lambda} \geq \mathbf{0}$. Moreover, let $\{\mathbf{T}_g\} \in \mathcal{S}_+^p$. Then if $\text{diag}(\mathbf{\Lambda}) > \mathbf{0}$, the problem of (5) is strictly concave.*

Proof of Lemma 3. By $\text{diag}(\mathbf{\Lambda}) > \mathbf{0}$, it is clear that the fused ridge penalty (4) is strictly convex as it is a conical combination of strictly convex and convex functions. Hence, the negative fused ridge penalty is strictly concave. The log-likelihood of (3) is a conical combination of concave functions and is thus also concave. Therefore, the penalized log-likelihood is strictly concave. \square

B.2. Proofs and additional results.

Proof of Proposition 1. To find the maximizing argument for a specific class of the general fused ridge penalized log-likelihood problem (5) we must obtain its first-order derivative w.r.t. that class and solve the resulting zero gradient equation. To this end we first rewrite the ridge penalty (4) into a second alternative form. Using that $\mathbf{\Lambda} = \mathbf{\Lambda}^\top$, and keeping in mind the cyclic property of the trace as well as properties of $\mathbf{\Omega}_g$ and \mathbf{T}_g stemming from their symmetry, we may find:

$$(23) \quad \begin{aligned} f^{\text{FR}''}(\{\mathbf{\Omega}_g\}; \mathbf{\Lambda}, \{\mathbf{T}_g\}) &= \sum_g \frac{\lambda_{gg}}{2} \|\mathbf{\Omega}_g - \mathbf{T}_g\|_F^2 + \sum_{g_1, g_2} \frac{\lambda_{g_1 g_2}}{4} \|(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1}) - (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})\|_F^2 \\ &= \sum_g \frac{\lambda_{g\bullet}}{2} \text{tr}[(\mathbf{\Omega}_g - \mathbf{T}_g)^\top (\mathbf{\Omega}_g - \mathbf{T}_g)] - \sum_{\substack{g_1, g_2 \\ g_1 \neq g_2}} \frac{\lambda_{g_1 g_2}}{2} \text{tr}[(\mathbf{\Omega}_{g_1} - \mathbf{T}_{g_1})^\top (\mathbf{\Omega}_{g_2} - \mathbf{T}_{g_2})], \end{aligned}$$

where $\lambda_{g\bullet} = \sum_{g'} \lambda_{gg'}$ denotes the sum over the g th row (or column) of $\mathbf{\Lambda}$. Taking the first-order partial derivative of (23) w.r.t. $\mathbf{\Omega}_{g_0}$ yields:

$$(24) \quad \begin{aligned} & \frac{\partial}{\partial \mathbf{\Omega}_{g_0}} f^{\text{FR}''}(\{\mathbf{\Omega}_g\}; \mathbf{\Lambda}, \{\mathbf{T}_g\}) \\ &= \lambda_{g_0\bullet} [2(\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) - (\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) \circ \mathbf{I}_p] - \sum_{g \neq g_0} \lambda_{gg_0} [2(\mathbf{\Omega}_g - \mathbf{T}_g) - (\mathbf{\Omega}_g - \mathbf{T}_g) \circ \mathbf{I}_p]. \end{aligned}$$

The first-order partial derivative of (3) w.r.t. $\mathbf{\Omega}_{g_0}$ results in:

$$(25) \quad \begin{aligned} \frac{\partial}{\partial \mathbf{\Omega}_{g_0}} \mathcal{L}(\{\mathbf{\Omega}_g\}; \{\mathbf{S}_g\}) &= \frac{\partial}{\partial \mathbf{\Omega}_{g_0}} \sum_g n_g \{ \ln |\mathbf{\Omega}_g| - \text{tr}(\mathbf{S}_g \mathbf{\Omega}_g) \}, \\ &= n_{g_0} [2(\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) - (\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) \circ \mathbf{I}_p]. \end{aligned}$$

Subtracting (24) from (25) yields

$$(26) \quad \left[n_{g_0} (\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) - \lambda_{g_0\bullet} (\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \lambda_{gg_0} (\mathbf{\Omega}_g - \mathbf{T}_g) \right] \circ (2\mathbf{J}_p - \mathbf{I}_p),$$

which, clearly, is $\mathbf{0}$ only when $n_{g_0} (\mathbf{\Omega}_{g_0}^{-1} - \mathbf{S}_{g_0}) - \lambda_{g_0\bullet} (\mathbf{\Omega}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \lambda_{gg_0} (\mathbf{\Omega}_g - \mathbf{T}_g) = \mathbf{0}$. From (26) we may then find our (conveniently scaled) zero gradient equation to be:

$$(27) \quad \hat{\mathbf{\Omega}}_{g_0}^{-1} - \mathbf{S}_{g_0} - \frac{\lambda_{g_0\bullet}}{n_{g_0}} (\hat{\mathbf{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}} (\mathbf{\Omega}_g - \mathbf{T}_g) = \mathbf{0}.$$

Now, rewrite (27) to

$$(28) \quad \hat{\mathbf{\Omega}}_{g_0}^{-1} - \bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0} (\hat{\mathbf{\Omega}}_{g_0} - \bar{\mathbf{T}}_{g_0}) = \mathbf{0},$$

where $\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} - \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}} (\mathbf{\Omega}_g - \mathbf{T}_g)$, $\bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0}$, and $\bar{\lambda}_{g_0} = \lambda_{g_0\bullet}/n_{g_0}$. It can be seen that (28) is of the form (22). Lemma 1 may then be applied to obtain the solution (7). \square

Corollary 1. Consider the estimator (7). Let $\hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g})$ be the precision matrix estimate of the g th class. Also, let $\text{diag}(\mathbf{\Lambda}) > \mathbf{0}$ and assume that all off-diagonal elements of $\mathbf{\Lambda}$ are zero. Then $\hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g})$ reduces to the non-fused ridge estimate of class g :

$$(29) \quad \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g}) = \hat{\mathbf{\Omega}}_g(\lambda_{gg}) = \left\{ \left[\frac{\lambda_{gg}}{n_g} \mathbf{I}_p + \frac{1}{4} \left(\mathbf{S}_g - \frac{\lambda_{gg}}{n_g} \mathbf{T}_g \right)^2 \right]^{1/2} + \frac{1}{2} \left(\mathbf{S}_g - \frac{\lambda_{gg}}{n_g} \mathbf{T}_g \right) \right\}^{-1}.$$

Proof of Corollary 1. The result follows directly from equations (7) and (8) by using that $\sum_{g' \neq g} \lambda_{gg'} = \sum_{g' \neq g} \lambda_{g'g} = 0$ for all g . \square

Lemma 4. Let $\{\mathbf{T}_g\} \in \mathcal{S}_+^p$ and assume $\lambda_{gg} \in \mathbb{R}_{++}$ in addition to $0 \leq \lambda_{gg'} < \infty$ for all $g' \neq g$. Then

$$\lim_{\lambda_{gg} \rightarrow \infty^-} \left\| \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g}) \right\|_F < \infty.$$

Proof of Lemma 4. The result is shown through proof by contradiction. Hence, suppose

$$\lim_{\lambda_{gg} \rightarrow \infty^-} \left\| \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g}) \right\|_F$$

is unbounded. Let $d[\cdot]_{jj}$ denote the j th largest eigenvalue. Then, as

$$\left\| \hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g}) \right\|_F = \left\{ \sum_{j=1}^p d \left[\hat{\mathbf{\Omega}}_g(\mathbf{\Lambda}, \{\mathbf{\Omega}_{g'}\}_{g' \neq g}) \right]_{jj}^2 \right\}^{1/2},$$

at least one eigenvalue must tend to infinity along with λ_{gg} . Assume without loss of generality that this is only the first (and largest) eigenvalue:

$$(30) \quad \lim_{\lambda_{gg} \rightarrow \infty^-} d \left[\hat{\Omega}_g(\Lambda, \{\Omega_{g'}\}_{g' \neq g}) \right]_{11} = \mathcal{O}(\lambda_{gg}^\gamma),$$

for some $\gamma > 0$. Now, for any λ_{gg} , the precision can be written as an eigendecomposition:

$$(31) \quad \hat{\Omega}_g(\Lambda, \{\Omega_{g'}\}_{g' \neq g}) = d_{11} \mathbf{v}_1 \mathbf{v}_1^\top + \sum_{j=2}^p d_{jj} \mathbf{v}_j \mathbf{v}_j^\top,$$

where the dependency of the eigenvalues and eigenvectors on the target matrices and penalty parameters has been suppressed (for notational brevity and clarity). It is the first summand on the right-hand side that dominates the precision for large λ_{gg} . Furthermore, this ridge ML precision estimate of the g th group satisfies, by (26), the following gradient equation:

$$n_g(\hat{\Omega}_g^{-1} - \mathbf{S}_g) - \lambda_{gg}(\hat{\Omega}_g - \mathbf{T}_g) - \sum_{g' \neq g} \lambda_{g'g}(\hat{\Omega}_g - \mathbf{T}_g) + \sum_{g' \neq g} \lambda_{g'g}(\Omega_{g'} - \mathbf{T}_{g'}) = \mathbf{0}.$$

We now make three observations: (i) From item (i) of Proposition 2 it follows that $\hat{\Omega}_g(\Lambda, \{\Omega_{g'}\}_{g' \neq g})$ is always p.d. for $\lambda_{gg} \in \mathbb{R}_{++}$. Consequently, $\lim_{\lambda_{gg} \rightarrow \infty^-} \|\hat{\Omega}_g(\Lambda, \{\Omega_{g'}\}_{g' \neq g})^{-1}\|_F < \infty$; (ii) The target matrices do not depend on λ_{gg} ; and (iii) The finite $\lambda_{gg'}$ ensure that the norms of $\Omega_{g'}$ can only exceed the norm of $\hat{\Omega}_g$ by a function (independent of λ_{gg}) of the constant $\lambda_{gg'}$. Hence, in the limit, the norms of the $\Omega_{g'}$ cannot exceed the norm of $\hat{\Omega}_g$. These observations give that, as λ_{gg} tends towards infinity, the term $\lambda_{gg}(\hat{\Omega}_g - \mathbf{T}_g)$ will dominate the gradient equation. In fact, the term $\lambda_{gg}\hat{\Omega}_g$ will dominate as, using (30) and (31):

$$\begin{aligned} \mathbf{0} &\approx -\lambda_{gg}(\hat{\Omega}_g - \mathbf{T}_g) \\ &\approx -\lambda_{gg}d_{11}\mathbf{v}_1\mathbf{v}_1^\top + \lambda_{gg}\mathbf{T}_g \\ &\approx -\lambda_{gg}^{1+\gamma}\mathbf{v}_1\mathbf{v}_1^\top + \lambda_{gg}\mathbf{T}_g \\ &\approx -\lambda_{gg}^{1+\gamma}(\mathbf{v}_1\mathbf{v}_1^\top + \lambda_{gg}^{-\gamma}\mathbf{T}_g) \\ &\approx -\lambda_{gg}^{1+\gamma}\mathbf{v}_1\mathbf{v}_1^\top. \end{aligned}$$

This latter statement is contradictory as it can only be true if the first eigenvalue tends to zero. This, in turn, contradicts the assumption of unboundedness (in the Frobenius norm) of the precision estimate. Hence, the fused ridge ML precision estimate must be bounded. \square

Proof of Proposition 2.

(i) Note that (27) for class g may be rewritten to

$$\hat{\Omega}_g^{-1} - \mathbf{S}_g - \frac{\lambda_{g\bullet}}{n_g} \left\{ \hat{\Omega}_g - \left[\mathbf{T}_g + \sum_{g' \neq g} \frac{\lambda_{gg'}}{\lambda_{g\bullet}} (\Omega_{g'} - \mathbf{T}_{g'}) \right] \right\} = \mathbf{0},$$

implying that (7) can be obtained under the following alternative updating scheme to (8):

$$\bar{\mathbf{S}}_g = \mathbf{S}_g, \quad \bar{\mathbf{T}}_g = \mathbf{T}_g + \sum_{g' \neq g} \frac{\lambda_{gg'}}{\lambda_{g\bullet}} (\Omega_{g'} - \mathbf{T}_{g'}), \quad \text{and} \quad \bar{\lambda}_g = \frac{\lambda_{g\bullet}}{n_g}.$$

Now, let $d[\cdot]_{jj}$ denote the j th largest eigenvalue. Then

$$d \left\{ [\hat{\Omega}_g]^{-1} \right\}_{jj} = d \left[\frac{1}{2} (\mathbf{S}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g) \right]_{jj} + \sqrt{\left\{ d \left[\frac{1}{2} (\mathbf{S}_g - \bar{\lambda}_g \bar{\mathbf{T}}_g) \right]_{jj} \right\}^2 + \bar{\lambda}_g} > 0,$$

when $\bar{\lambda}_g > 0$. As $\bar{\lambda}_g = \sum_{g'} (\lambda_{g'g}/n_g)$ and as $\lambda_{g'g}$ may be 0 for all $g' \neq g$, $\hat{\Omega}_g$ is guaranteed to be p.d. whenever $\lambda_{gg} \in \mathbb{R}_{++}$.

(ii) Note that $\sum_{g' \neq g} \lambda_{gg'} = \sum_{g' \neq g} \lambda_{g'g} = 0$ implies that $\hat{\Omega}_g$ reduces to the non-fused class estimate (29) by way of Corollary 1. The stated right-hand limit is then immediate by using $\lambda_{gg} = 0$ in (29). Under the distributional assumptions this limit exists with probability 1 when $p \leq n_g$.

(iii) Consider the zero gradient equation (27) for the g th class. Multiply it by $n_g/\lambda_{g\bullet}$ to factor out the dominant term:

$$(32) \quad \frac{n_g}{\lambda_{g\bullet}} \hat{\Omega}_g^{-1} - \frac{n_g}{\lambda_{g\bullet}} \mathbf{S}_g - (\hat{\Omega}_g - \mathbf{T}_g) + \sum_{g' \neq g} \frac{\lambda_{g'g}}{\lambda_{g\bullet}} (\Omega_{g'} - \mathbf{T}_{g'}) = \mathbf{0}.$$

When $\lambda_{gg} \rightarrow \infty^-$, $\lambda_{g\bullet} = \sum_{g'} \lambda_{gg'} \rightarrow \infty^-$, implying that the first two terms of (32) vanish. Under the assumption that $\lambda_{gg'} < \infty$ for all $g' \neq g$ we have that $\lambda_{g'g}/\lambda_{g\bullet} \rightarrow 0$ when $\lambda_{gg} \rightarrow \infty^-$ for all $g' \neq g$. Thus, all terms of the sum also vanish as Lemma 4 implies that the $\Omega_{g'}$ are all bounded. Hence, when $\lambda_{gg} \rightarrow \infty^-$ and $\lambda_{gg'} < \infty$ for all $g' \neq g$, the zero gradient equation reduces to $\hat{\Omega}_g - \mathbf{T}_g = \mathbf{0}$, implying the stated left-hand limit.

(iv) The proof strategy follows the proof of item iii. Multiply the zero gradient equation (27) for the g_1 th class with $n_{g_1}/\lambda_{g_1g_2}$ to obtain:

$$(33) \quad \frac{n_{g_1}}{\lambda_{g_1g_2}} \hat{\Omega}_{g_1}^{-1} - \frac{n_{g_1}}{\lambda_{g_1g_2}} \mathbf{S}_{g_1} - \frac{\lambda_{g_1\bullet}}{\lambda_{g_1g_2}} (\hat{\Omega}_{g_1} - \mathbf{T}_{g_1}) + \sum_{g' \neq g_1} \frac{\lambda_{g'g_1}}{\lambda_{g_1g_2}} (\Omega_{g'} - \mathbf{T}_{g'}) = \mathbf{0}.$$

The first two terms are immediately seen to vanish when $\lambda_{g_1g_2} \rightarrow \infty^-$. Under the assumption that all penalties except $\lambda_{g_1g_2}$ are finite, we have that $\lambda_{g_1\bullet}/\lambda_{g_1g_2} \rightarrow 1$ for $\lambda_{g_1g_2} \rightarrow \infty^-$. Similarly, all elements of the sum term in (33) vanish except the element where $g' = g_2$. Hence, when $\lambda_{g_1g_2} \rightarrow \infty^-$ and when $\lambda_{g'_1g'_2} < \infty$ for all $\{g'_1, g'_2\} \neq \{g_1, g_2\}$, the zero gradient equation for class g_1 reduces to:

$$(34) \quad -(\hat{\Omega}_{g_1} - \mathbf{T}_{g_1}) + (\Omega_{g_2} - \mathbf{T}_{g_2}) = \mathbf{0}.$$

Conversely, by multiplying the zero gradient equation (27) for the g_2 th class with $n_{g_2}/\lambda_{g_1g_2}$ one obtains, through the same development as above, that the zero gradient equation for class g_2 reduces to the $\hat{\Omega}_{g_2}$ -analogy of equation (34). The result (34) then immediately implies the stated limiting result. \square

Corollary 2. Consider item iv of Proposition 2. When, in addition, $\mathbf{T}_{g_1} = \mathbf{T}_{g_2}$, we have that

$$\lim_{\lambda_{g_1g_2} \rightarrow \infty^-} (\hat{\Omega}_{g_1} - \mathbf{T}_{g_1}) = \lim_{\lambda_{g_1g_2} \rightarrow \infty^-} (\hat{\Omega}_{g_2} - \mathbf{T}_{g_2}) \implies \hat{\Omega}_{g_1} = \hat{\Omega}_{g_2}.$$

Proof of Corollary 2. The implication follows directly by using $\mathbf{T}_{g_1} = \mathbf{T}_{g_2}$ in (34). \square

Proof of Proposition 3. The result follows directly from Proposition 1 and Lemma 2. \square

Proof of Proposition 4. Note that line 8 of Algorithm 1 implies that the initializing estimates are p.d. Moreover, regardless of the value of the fused penalties (in the feasible domain), the estimate in line 11 of Algorithm 1 is p.d. as a consequence of Proposition 2. \square

REFERENCES

- [1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.
- [2] O. Banerjee, L. El Ghaoui, and A. D’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9: 485–516, 2008.
- [3] A. L. Barabási. Scale-free networks: A decade and beyond. *Science*, 325(5939):412–413, 2009.

- [4] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [5] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.
- [6] A. K. Bera and Y. Biliyas. Rao’s score, Neyman’s $c(\alpha)$ and Silvey’s LM tests: An essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, 97(1):9–44, 2001.
- [7] N. Bidère, V. N. Ngo, J. Lee, C. Collins, L. Zheng, F. Wan, R. E. Davis, G. Lenz, D. E. Anderson, D. Arnoult, A. Vazquez, K. Sakai, J. Zhang, Z. Meng, T. D. Veenstra, L. M. Staudt, and M. J. Lenardo. Casein kinase 1 α governs antigen-receptor-induced NF- κ B activation and human lymphoma cell survival. *Nature*, 458(7234):92–96, 2009.
- [8] A. E. Bilgrau and S. Falgreen. *DLBCLdata: Automated and Reproducible Download and Preprocessing of DLBCL Data*, 2014. URL <http://github.com/AEBilgrau/DLBCLdata>. R package version 0.9.
- [9] A. E. Bilgrau, P. S. Eriksen, K. Dybkær, and M. Bøgstød. Estimation of a common covariance matrix for multiple classes with applications in meta- and discriminant analysis. *Submitted to Annals of Applied Statistics*, *arXiv:1503.07990*, 269553, 2015.
- [10] J. L. Browning, I. D. Sizing, P. Lawton, P. R. Bourdon, P. D. Rennert, G. R. Majeau, C. M. Ambrose, C. Hession, K. Miatkowski, D. A. Griffiths, Ngam ek A., Meier W., Benjamin C. D., and Hochman P. S. Characterization of lymphotoxin- $\alpha\beta$ complexes on the surface of mouse lymphocytes. *The Journal of Immunology*, 159(7):3288–3298, 1997.
- [11] M. A. Care, S. Barrans, L. Worrillow, A. Jack, D. R. Westhead, and R. M. Tooze. A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PLoS One*, 8(2):e55895, 2013.
- [12] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33(20):e175, 2005.
- [13] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76(2):373–397, 2014.
- [14] K. Dybkær, M. Bøgstød, S. Falgreen, J. S. Bødker, M. K. Kjeldsen, A. Schmitz, A. E. Bilgrau, Z. Y. Xu-Monette, L. Li, K. S. Bergkvist, M. B. Laursen, M. Rodrigo-Domingo, S. C. Marques, S. B. Rasmussen, M. Nyegaard, M. Gaihede, M. B. Møller, R. J. Samworth, R. D. Shah, P. Johansen, T. C. El-Galaly, K. H. Young, and H. E. Johnsen. A diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *Journal Of Clinical Oncology*, 33(12):1379–1388, 2015.
- [15] D. Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer-Verlag, New York, 2013.
- [16] D. Eddelbuettel and R. François. *Rcpp: Seamless R and C++ integration*. *Journal of Statistical Software*, 40(8), 2011.
- [17] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.
- [18] R. François, D. Eddelbuettel, and D. Bates. *RcppArmadillo: Rcpp Integration for Armadillo Templated Linear Algebra Library*, 2012. URL <http://CRAN.R-project.org/package=RcppArmadillo>. R package version 0.3.6.1.
- [19] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–41, 2008.
- [20] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. *affy*—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.
- [21] Y. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.
- [22] B. Jones and M. West. Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, 92:779–786, 2005.

- [23] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
- [24] S. L. Lauritzen. *Graphical models*. Clarendon Press, Oxford, 1996.
- [25] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc., 2010.
- [26] X. Lu and X. Zhang. The effect of GeneChip gene definitions on the microarray study of cancers. *Bioessays*, 28(7):739–46, 2006.
- [27] S. Mei, X. Zhang, and M. Cao. *Power Grid Complexity*. Tsinghua University Press, Beijing and Springer-Verlag Berlin, 2011.
- [28] O. Mersmann. *microbenchmark: Accurate Timing Functions*, 2014. URL <http://CRAN.R-project.org/package=microbenchmark>. R package version 1.4-2.
- [29] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, 2010.
- [30] G. S. Nowakowski, B. LaPlant, W. R. Macon, C. B. Reeder, J. M. Foran, G. D. Nelson, C. A. Thompson, C. E. Rivera, D. J. Inwards, I. N. Micallef, P. B. Johnston, L. F. Porrata, S. M. Ansell, R. D. Gascoyne, T. M. Habermann, and T. E. Witzig. Lenalidomide combined with R-CHOP overcomes negative prognostic impact of non-germinal center B-cell phenotype in newly diagnosed diffuse large B-cell lymphoma: A phase II study. *Journal of Clinical Oncology*, 33(3):251–257, 2015.
- [31] C. F. W. Peeters, A. E. Bilgrau, and W. N. van Wieringen. *rags2ridges: Ridge Estimation of Precision Matrices from High-Dimensional Data*, forthcoming. R package version 2.0.
- [32] C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.
- [33] B. S. Price, C. J. Geyer, and A. J. Rothman. Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics*, 24(2):439–454, 2015.
- [34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL <http://www.R-project.org/>.
- [35] M. Roschewski, L. M. Staudt, and W. H. Wilson. Diffuse large B-cell lymphoma-treatment approaches in the molecular era. *Nature Reviews Clinical Oncology*, 11(1):12–23, 2014.
- [36] J. Ruan, P. Martin, R. R. Furman, S. M. Lee, K. Cheung, J. M. Vose, A. LaCasce, J. Morrison, R. Elstrom, S. Ely, A. Chadburn, E. Cesarman, M. Coleman, and J. P. Leonard. Bortezomib plus CHOP-rituximab for previously untreated diffuse large B-cell lymphoma and mantle cell lymphoma. *Journal of Clinical Oncology*, 29(6):690–697, 2011.
- [37] R. Sandberg and O. Larsson. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8(1):48, 2007.
- [38] C. Sanderson. *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Technical Report, NICTA, 2010. URL <http://arma.sourceforge.net>.
- [39] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:art. 32, 2005.
- [40] J. M. Schuetz, N. A. Johnson, R. D. Morin, D. W. Scott, K. Tan, S Ben-Nierah, M Boyle, G. W. Slack, M. A. Marra, J. M. Connors, A. R. Brooks-Wilson, and R. D. Gascoyne. BCL2 mutations in diffuse large B-cell lymphoma. *Leukemia*, 26(6):1383–90, 2012.
- [41] The Non-Hodgkin’s Lymphoma Classification Project. A clinical evaluation of the international lymphoma study group classification of non-Hodgkin’s lymphoma. *Blood*, 89(11):3909–3918, 1997.
- [42] W. N. van Wieringen and C. F. W. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. Submitted to *Computational Statistics & Data Analysis*, *arXiv:1403.0904v3*, 2015.
- [43] I. Vujčić, A. Abbruzzo, and E. Wit. A computationally fast alternative to cross-validation in penalized gaussian graphical models. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–13,

2015. doi: 10.1080/00949655.2014.992020.
- [44] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684): 440–442, 1998.
 - [45] L. Williams-Abbott, B. N. Walter, T. C. Cheung, C. R. Goh, A. G. Porter, and C. F. Ware. The lymphotoxin- α ($lt\alpha$) subunit is essential for the assembly, but not for the receptor specificity, of the membrane-anchored $lt\alpha 1\beta 2$ heterotrimeric ligand. *The Journal of Biological Chemistry*, 271(31):19451–6, 1997.
 - [46] D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society, Series B*, 71:615–636, 2009.
 - [47] Y. Yang, A. L. Shaffer, N. C. T. Emre, M. Ceribelli, M. Zhang, G. Wright, W. Xiao, J. Powell, J. Platig, H. Kohlhammer, Young R. M., H. Zhao, Y. Yang, W. Xu, J. J. Buggy, S. Balasubramanian, L. A. Mathews, P. Shinn, R. Guha, M. Ferrer, C. Thomas, T. A. Waldmann, and L. M. Staudt. Exploiting synthetic lethality for the therapy of ABC diffuse large B cell lymphoma. *Cancer cell*, 21(6):723–737, 2012.
 - [48] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94: 19–35, 2007.
 - [49] Y. Yuan. Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17:809–826, 2008.
 - [50] J. D. Zhang and S. Wiemann. **KEGGgraph**: A graph approach to KEGG pathway in R and Bioconductor. *Bioinformatics*, 25(11):1470–1471, 2009.

(Anders Ellern Bilgrau) DEPT. OF MATHEMATICAL SCIENCES, AALBORG UNIVERSITY, AALBORG, DENMARK; AND DEPT. OF HAEMATOLOGY, AALBORG UNIVERSITY HOSPITAL, AALBORG, DENMARK

E-mail address: `abilgrau@math.aau.dk`

(Carel F.W. Peeters) DEPT. OF EPIDEMIOLOGY & BIostatISTICS, VU UNIVERSITY MEDICAL CENTER AMSTERDAM, AMSTERDAM, THE NETHERLANDS

E-mail address: `cf.peeters@vumc.nl`

(Poul Svante Eriksen) DEPT. OF MATHEMATICAL SCIENCES, AALBORG UNIVERSITY, AALBORG, DENMARK

E-mail address: `svante@math.aau.dk`

(Martin Bøgsteds) DEPT. OF HAEMATOLOGY, AALBORG UNIVERSITY HOSPITAL, AALBORG, DENMARK

E-mail address: `m.boegsted@dcu.aau.dk`

(Wessel N. van Wieringen) DEPT. OF EPIDEMIOLOGY & BIostatISTICS, VU UNIVERSITY MEDICAL CENTER AMSTERDAM, AMSTERDAM, THE NETHERLANDS; AND DEPT. OF MATHEMATICS, VU UNIVERSITY AMSTERDAM, AMSTERDAM, THE NETHERLANDS

E-mail address: `w.vanwieringen@vumc.nl`

SUPPLEMENTARY MATERIAL

1. ALTERNATIVE FUSED RIDGE SOLUTIONS

This section derives two equivalent (in terms of (7)) alternative updating schemes to (8). The motivation for the exploration of these alternative recursive estimators is twofold. First, alternative recursions can exhibit differing numerical (in)stability for extreme values of the penalty matrix $\mathbf{\Lambda} = [\lambda_{g_1 g_2}]$. Second, they provide additional intuition and understanding of the targeted fused ridge estimator.

The general strategy to finding the alternatives is to rewrite the gradient equation (27) into the non-fused form (28), which we will repeat here:

$$(S1) \quad \hat{\mathbf{\Omega}}_{g_0}^{-1} - \bar{\mathbf{S}}_{g_0} - \bar{\lambda}_{g_0}(\hat{\mathbf{\Omega}}_{g_0} - \bar{\mathbf{T}}_{g_0}) = \mathbf{0},$$

where $\bar{\lambda}_{g_0}$, $\bar{\mathbf{T}}_{g_0}$, and $\bar{\mathbf{S}}_{g_0}$ do not depend on $\hat{\mathbf{\Omega}}_{g_0}$. Note that an explicit closed-form solution to (S1) exists in the form of (7).

1.1. First alternative. The first alternative scheme is straightforward. Rewrite (27) to:

$$(S2) \quad \begin{aligned} \mathbf{0} &= n_{g_0} \hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0} \mathbf{S}_{g_0} - \lambda_{g_0 \bullet} (\hat{\mathbf{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \sum_{g \neq g_0} \lambda_{gg_0} (\mathbf{\Omega}_g - \mathbf{T}_g) \\ &= n_{g_0} \hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0} \mathbf{S}_{g_0} - \lambda_{g_0 \bullet} \left\{ \hat{\mathbf{\Omega}}_{g_0} - \left[\mathbf{T}_{g_0} + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{\lambda_{g_0 \bullet}} (\mathbf{\Omega}_g - \mathbf{T}_g) \right] \right\}, \end{aligned}$$

where $\lambda_{g_0 \bullet} = \sum_g \lambda_{gg_0}$. In terms of (S1), we thus have the updating scheme given in equation (9). As stated in the main text, it has the intuitive interpretation that a fused class target is used which is a combination of the class-specific target and the ‘target corrected’ estimates of remaining classes.

1.2. Second alternative. We now derive a second alternative recursion scheme. Add and subtract $\lambda_{g_0 \bullet} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g$, to (S2) and rewrite such that:

$$\begin{aligned} \mathbf{0} &= n_{g_0} \hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0} \mathbf{S}_{g_0} - \lambda_{g_0 \bullet} (\hat{\mathbf{\Omega}}_{g_0} - \mathbf{T}_{g_0}) + \lambda_{g_0 \bullet} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g + \sum_{g \neq g_0} \lambda_{gg_0} (\mathbf{\Omega}_g - \mathbf{T}_g) - \lambda_{g_0 \bullet} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \\ &= n_{g_0} \hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0} \mathbf{S}_{g_0} - \lambda_{g_0 \bullet} \left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \right) \right] + \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g - \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{T}_g - \lambda_{g_0 \bullet} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \\ &= n_{g_0} \hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0} \mathbf{S}_{g_0} - \lambda_{g_0 \bullet} \left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \right) \right] - \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{T}_g - (\lambda_{g_0 \bullet} - 1) \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \\ &= n_{g_0} \hat{\mathbf{\Omega}}_{g_0}^{-1} - n_{g_0} \left[\mathbf{S}_{g_0} + \frac{\lambda_{g_0 \bullet} - 1}{n_{g_0}} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}} \mathbf{T}_g \right] - \lambda_{g_0 \bullet} \left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \right) \right]. \end{aligned}$$

Dividing by n_{g_0} gives

$$\mathbf{0} = \hat{\mathbf{\Omega}}_{g_0}^{-1} - \left[\mathbf{S}_{g_0} + \frac{\lambda_{g_0 \bullet} - 1}{n_{g_0}} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}} \mathbf{T}_g \right] - \frac{\lambda_{g_0 \bullet}}{n_{g_0}} \left[\hat{\mathbf{\Omega}}_{g_0} - \left(\mathbf{T}_{g_0} + \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g \right) \right],$$

which brings the expression to the desired form (S1) with the updating scheme

$$\bar{\mathbf{S}}_{g_0} = \mathbf{S}_{g_0} + \frac{\lambda_{g_0 \bullet} - 1}{n_{g_0}} \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g + \sum_{g \neq g_0} \frac{\lambda_{gg_0}}{n_{g_0}} \mathbf{T}_g, \quad \bar{\mathbf{T}}_{g_0} = \mathbf{T}_{g_0} + \sum_{g \neq g_0} \lambda_{gg_0} \mathbf{\Omega}_g, \quad \text{and} \quad \bar{\lambda}_{g_0} = \frac{\lambda_{g_0 \bullet}}{n_{g_0}}.$$

Again, a solution for $\hat{\mathbf{\Omega}}_{g_0}$ with fixed $\mathbf{\Omega}_g$ for all $g \neq g_0$, is available through Lemma 1 [42] and is given in (7).

1.3. Motivation. Though seemingly more complicated, these alternative updating schemes can be numerically more stable for extreme penalties. In both alternatives, we see that $\hat{\mathbf{S}}_{g_0}$ is p.s.d. for (nearly) all very large and very small penalties. Likewise, $\hat{\mathbf{T}}_{g_0}$ is always positive definite. Compare the alternative expressions to the updating scheme given by (8) which can be seen to be numerically unstable for very large penalties: For very large λ_{gg} or $\lambda_{g_1g_2}$ the $\hat{\mathbf{S}}_{g_0}$ in (8) may be a matrix with numerically extreme values. This implies ill-conditioning and numerical instability under finite computer precision. On the other hand, ‘updating’ the target matrix will generally lead to updates for which the resulting estimator is not rotationally equivariant. This implies a reduction in computational speed.

2. ESTIMATION IN SPECIAL CASES

Here we explore scenarios for which we arrive at explicit targeted fused ridge estimators. These explicit solutions further insight into the behavior of the general estimator and they can provide computational speed-ups in certain situations. Three special cases are covered:

- I. $\lambda_{gg'} = 0$ for all $g \neq g'$ or equivalently $\sum_{g'} \lambda_{gg'} = \lambda_{g\bullet} = \lambda_{gg}$ for all g ;
- II. $\mathbf{\Omega}_1 = \dots = \mathbf{\Omega}_G$ and $\mathbf{T}_g = \mathbf{T}$ for all g ;
- III. $\mathbf{T}_g = \mathbf{T}$ for all g , $\lambda_{gg} = \lambda$ for all g , $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$, and $\lambda_f \rightarrow \infty^-$.

2.1. Special case I. When $\sum_{g'} \lambda_{gg'} = \lambda_{g\bullet} = \lambda_{gg}$ for all g , we have that $\sum_{g' \neq g} \lambda_{gg'} = \sum_{g' \neq g} \lambda_{g'g} = 0$ for all g . Hence, all fusion penalties are zero. The zero gradient equation (27) for class g then no longer hinges upon information from the remaining classes g' . The targeted fused precision estimate for class g then reduces to (29) of Corollary 1. This case thus coincides, as expected, with obtaining G decoupled non-fused ridge precision estimates. A special case that results in the same estimates occurs when considering $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$ and λ_f is taken to be 0.

2.2. Special case II. Suppose $\mathbf{\Omega}_g = \mathbf{\Omega}$ and $\mathbf{T}_g = \mathbf{T}$ for all g . Consequently, the fusion penalty term vanishes irrespective of the values of the $\lambda_{g_1g_2}$, $g_1 \neq g_2$. The zero gradient equation (27) then reduces to

$$\mathbf{0} = n_g \hat{\mathbf{\Omega}}^{-1} - n_g \mathbf{S}_g - \lambda_{gg} (\hat{\mathbf{\Omega}} - \mathbf{T}),$$

for each class g . Adding all G equations implies:

$$\begin{aligned} \mathbf{0} &= \sum_{g=1}^G n_g \hat{\mathbf{\Omega}}^{-1} - \sum_{g=1}^G n_g \mathbf{S}_g - \left(\sum_{g=1}^G \lambda_{gg} \right) (\hat{\mathbf{\Omega}} - \mathbf{T}) \\ &= n_{\bullet} \hat{\mathbf{\Omega}}^{-1} - n_{\bullet} \mathbf{S}_{\bullet} - \text{tr}(\mathbf{\Lambda}) (\hat{\mathbf{\Omega}} - \mathbf{T}) \\ (S3) \quad &= \hat{\mathbf{\Omega}}^{-1} - \left[\mathbf{S}_{\bullet} - \frac{\text{tr}(\mathbf{\Lambda})}{n_{\bullet}} \mathbf{T} \right] - \frac{\text{tr}(\mathbf{\Lambda})}{n_{\bullet}} \hat{\mathbf{\Omega}}. \end{aligned}$$

We recognize that (S3) is of the form (22). Lemma 1 may then be directly applied to obtain the solution:

$$(S4) \quad \hat{\mathbf{\Omega}}(\mathbf{\Lambda}) = \left\{ \left[\lambda^* \mathbf{I}_p + \frac{1}{4} (\mathbf{S}_{\bullet} - \lambda^* \mathbf{T})^2 \right]^{1/2} + \frac{1}{2} (\mathbf{S}_{\bullet} - \lambda^* \mathbf{T}) \right\}^{-1},$$

where $\lambda^* = \text{tr}(\mathbf{\Lambda})/n_{\bullet}$. Hence, this second special case gives a non-fused penalized estimate that uses the pooled covariance matrix. It can be interpreted as an averaged penalized estimator. It is of importance in testing equality of the class precision matrices (see Section 4.1 of the main text).

2.3. Special case III. Suppose that $\mathbf{T}_g = \mathbf{T}$ for all g , that $\lambda_{gg} = \lambda$ for all g , and that $\lambda_{g_1g_2} = \lambda_f$ for all $g_1 \neq g_2$. The main optimization problem then reduces to (6). Clearly, for $\lambda_f \rightarrow \infty^-$ the fused penalty

$$f^{\text{FR}}(\{\mathbf{\Omega}_g\}; \lambda, \lambda_f, \mathbf{T}) = \frac{\lambda}{2} \sum_g \|\mathbf{\Omega}_g - \mathbf{T}\|_F^2 + \frac{\lambda_f}{4} \sum_{g_1, g_2} \|(\mathbf{\Omega}_{g_1} - \mathbf{\Omega}_{g_2})\|_F^2$$

is minimized when $\mathbf{\Omega}_1 = \mathbf{\Omega}_2 = \dots = \mathbf{\Omega}_G$. This is also implied, more rigorously, by Corollary 2. Hence, the problem reduces to the special case of section 2.2 considered above. The solution to the penalized ML problem when $\lambda_f = \infty$ is then given by (S4) where $\text{tr}(\mathbf{\Lambda})$ now implies $G\lambda$.

3. FUSED KULLBACK-LEIBLER APPROXIMATE CROSS-VALIDATION

3.1. Motivation. In ℓ_1 -penalized estimation of the precision matrix, penalty selection implies (graphical) model selection: Regularization results in automatic selection of conditional dependencies. One then seeks to select an optimal value for the penalty parameter in terms of model selection consistency. To this end, the Bayesian information criterion (BIC), the extended BIC (EBIC), and the stability approach to regularization selection (StARS) are appropriate [25]. The (fused) ℓ_2 -penalty will not directly induce sparsity in precision matrix estimates. Hence, in ℓ_2 -penalized problems it is natural to choose the penalty parameters on the basis of efficiency loss. Of interest are then estimators of the Kullback-Leibler (KL) divergence, such as LOOCV, generalized approximate cross-validation (GACV), and Akaike's information criterion (AIC). While superior in terms of predictive accuracy due to its data-driven nature, the LOOCV is computationally very expensive. Vujačić et al. [43] proposed a KL-based CV loss with superior performance to both AIC and GACV. The proposed method has closed-form solutions and thus provides a fast approximation to LOOCV. Here, we extend this method to provide a computationally friendly approximation of the fused LOOCV score.

3.2. Formulation. Following Vujačić et al. [43], we now restate the KL approximation to LOOCV in the fused ridge setting. Let the true precision matrix for class g be denoted by $\mathbf{\Omega}_g$. Its estimate, shorthand by $\hat{\mathbf{\Omega}}_g$ can be obtained through Algorithm 1. The KL divergence between the multivariate normal distributions $\mathcal{N}_p(\mathbf{0}, \mathbf{\Omega}_g^{-1})$ and $\mathcal{N}_p(\mathbf{0}, \hat{\mathbf{\Omega}}_g^{-1})$ can be shown to be:

$$\text{KL}(\mathbf{\Omega}_g, \hat{\mathbf{\Omega}}_g) = \frac{1}{2} \left\{ \text{tr}(\mathbf{\Omega}_g^{-1} \hat{\mathbf{\Omega}}_g) - \ln |\mathbf{\Omega}_g^{-1} \hat{\mathbf{\Omega}}_g| - p \right\}.$$

For each g we wish to minimize this divergence. In the fused case we therefore consider the *fused Kullback-Leibler* (FKL) divergence which, motivated by the LOOCV score, is taken to be a weighted average of KL divergences:

$$\begin{aligned} & \text{FKL}(\{\mathbf{\Omega}_g\}, \{\hat{\mathbf{\Omega}}_g\}) \\ (S5) \quad &= \frac{1}{n_{\bullet}} \sum_{g=1}^G n_g \text{KL}(\mathbf{\Omega}_g, \hat{\mathbf{\Omega}}_g) = \frac{1}{n_{\bullet}} \sum_{g=1}^G \frac{n_g}{2} \left\{ \text{tr}(\mathbf{\Omega}_g^{-1} \hat{\mathbf{\Omega}}_g) - \ln |\mathbf{\Omega}_g^{-1} \hat{\mathbf{\Omega}}_g| - p \right\}. \end{aligned}$$

The FKL divergence (S5) can, using the likelihood (3), be rewritten as

$$\text{FKL} = -\frac{1}{n_{\bullet}} \mathcal{L}(\{\hat{\mathbf{\Omega}}_g\}; \{\mathbf{S}_g\}) + \text{bias}, \quad \text{where} \quad \text{bias} = \frac{1}{2n_{\bullet}} \sum_{g=1}^G n_g \text{tr}[\hat{\mathbf{\Omega}}_g(\mathbf{\Omega}_g^{-1} - \mathbf{S}_g)],$$

and where the equality holds up to the addition of a constant. It is clear that the bias term depends on the unknown true precision matrices and thus needs to be estimated. The fused analogue to the proposal of Vujačić et al. [43], called the *fused Kullback-Leibler approximate cross-validation* score or simply *approximate fused LOOCV* score, then is

$$(S6) \quad \widehat{\text{FKL}}(\mathbf{\Lambda}) = -\frac{1}{n_{\bullet}} \mathcal{L}(\{\hat{\mathbf{\Omega}}_g\}; \{\mathbf{S}_g\}) + \widehat{\text{bias}},$$

with

$$(S7) \quad \widehat{\text{bias}} = \frac{1}{2n_{\bullet}} \sum_{g=1}^G \sum_{i=1}^{n_g} \left\{ \mathbf{y}_{ig}^{\top} (\hat{\mathbf{\Omega}}_g^2 - \hat{\mathbf{\Omega}}_g) \mathbf{y}_{ig} + \bar{\lambda}_g \mathbf{y}_{ig}^{\top} (\hat{\mathbf{\Omega}}_g^4 - \hat{\mathbf{\Omega}}_g^3) \mathbf{y}_{ig} \right\},$$

and where $\bar{\lambda}_g = \frac{\lambda_{g\bullet}}{n_g}$. The derivation of this estimate is given in Section B.2 below. One would then choose $\mathbf{\Lambda}^*$ such that the FKL approximate cross-validation score is minimized:

$$(S8) \quad \mathbf{\Lambda}^* = \arg \min_{\mathbf{\Lambda}} \widehat{\text{FKL}}(\mathbf{\Lambda}), \quad \text{subject to: } \mathbf{\Lambda} \geq \mathbf{0} \wedge \text{diag}(\mathbf{\Lambda}) > \mathbf{0}.$$

The closed form expression in (S6) implies that $\mathbf{\Lambda}^*$ is more rapidly determined than $\mathbf{\Lambda}^*$. As seen in the derivation, $\mathbf{\Lambda}^* \approx \mathbf{\Lambda}^*$ for large sample sizes.

3.3. Derivation. Here we give, borrowing some ideas from Vujačić et al. [43], the derivation of the estimate (S6). Let observation i in class g be denoted by \mathbf{y}_{ig} and let $\mathbf{S} = \mathbf{S}_{ig} = \mathbf{y}_{ig}\mathbf{y}_{ig}^\top$ be the sample covariance or scatter matrix of that observation. As before, the singularly indexed $\mathbf{S}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{S}_{ig}$ is the class-specific sample covariance matrix. Throughout this section we will conveniently drop (some of) the explicit notation.

The FKL divergence reframes the LOOCV score in terms of a likelihood evaluation and a bias term when \mathbf{S} is *not* left out of class g . We thus study the change in the estimate as function of the single scatter matrix \mathbf{S} . Let $\hat{\Omega}_g(\mathbf{S}) = \hat{\Omega}_g^{-ig}$ be the estimate in class g when \mathbf{S} is omitted. That is, $\hat{\Omega}_g(\mathbf{S})$ is part of the solution to the system

$$(S9) \quad \hat{\Omega}_a^{-1} + \mu_{aa}\hat{\Omega}_a + \mathbb{1}[a=g]\mathbf{S} + \sum_{b \neq a} \mu_{ab}\hat{\Omega}_b + \mathbf{A}_a = \mathbf{0}, \quad \text{for all } a = 1, \dots, G,$$

where $\mu_{aa} = -\frac{\lambda_{a\bullet}}{n_a}$, $\mu_{ab} = \frac{\lambda_{ab}}{n_a}$, and where \mathbf{A}_a is a matrix determined by the remaining data, penalty parameters and targets. Note that the penalized MLE can be denoted $\hat{\Omega}_g = \hat{\Omega}_g(\mathbf{0})$, which corresponds to the ‘full’ estimate resulting from the full gradient equation (27).

We wish to approximate $\hat{\Omega}_g(\mathbf{S})$ by a Taylor expansion around $\hat{\Omega}_g(\mathbf{0})$, i.e.:

$$\hat{\Omega}_a(\mathbf{S}) \approx \hat{\Omega}_a(\mathbf{0}) + \sum_{j,j'} \frac{\partial \hat{\Omega}_a}{\partial S_{jj'}} S_{jj'}.$$

Differentiating (S9) w.r.t. $S_{jj'}$, the (j, j') th entry in \mathbf{S} , and equating to zero yields

$$(S10) \quad \begin{aligned} \mathbf{0} &= -\hat{\Omega}_a^{-1} \frac{\partial \hat{\Omega}_a}{\partial S_{jj'}} \hat{\Omega}_a^{-1} + \mu_{aa} \frac{\partial \hat{\Omega}_a}{\partial S_{jj'}} + \mathbb{1}[a=g] \mathbf{E}_{jj'} + \sum_{b \neq a} \mu_{ab} \frac{\partial \hat{\Omega}_b}{\partial S_{jj'}} \\ &= -\hat{\Omega}_a^{-1} \frac{\partial \hat{\Omega}_a}{\partial S_{jj'}} \hat{\Omega}_a^{-1} + \sum_b \mu_{ab} \frac{\partial \hat{\Omega}_b}{\partial S_{jj'}} + \mathbb{1}[a=g] \mathbf{E}_{jj'}, \quad \text{for all } j, j', \end{aligned}$$

where $\mathbf{E}_{jj'}$ is the null matrix except for unity in entries (j, j') and (j', j) . The third term is obtained as $\partial \mathbf{S} / \partial S_{jj'} = \mathbf{E}_{jj'}$ by the symmetric structure of \mathbf{S} . This is also seen from the fact that $\mathbf{S} = \sum_{jj'} S_{jj'} \mathbf{E}_{jj'}$. Let

$$\mathbf{V}(\mathbf{S})_a = \sum_{j,j'} \frac{\partial \hat{\Omega}_a}{\partial S_{jj'}} S_{jj'},$$

and multiply (S10) by $S_{jj'}$ and sum over all j, j' to obtain

$$(S11) \quad \hat{\Omega}_a^{-1} \mathbf{V}(\mathbf{S})_a \hat{\Omega}_a^{-1} - \sum_b \mu_{ab} \mathbf{V}(\mathbf{S})_b = \mathbb{1}[a=g] \mathbf{S}, \quad \text{for all } a = 1, \dots, G.$$

We seek the solution vector $\mathbf{V} = \{\mathbf{V}(\mathbf{S})_a\}_{a=1}^G$ of square matrices for the system of equations in (S11) which can be rewritten in the following way. Introduce and consider the linear operator (or block matrix):

$$\mathbf{N} = \{\mathbf{N}_{ab}\}_{a,b=1}^G \quad \text{where} \quad \mathbf{N}_{ab} = \begin{cases} \hat{\Omega}_a^{-1} \otimes \hat{\Omega}_a^{-1} - \mu_{aa} \mathbf{I}_p \otimes \mathbf{I}_p & \text{if } a = b \\ -\mu_{ab} \mathbf{I}_p \otimes \mathbf{I}_p & \text{if } a \neq b \end{cases}.$$

Then \mathbf{V} can be verified to be the solution to the system (S10) as

$$\begin{aligned} \mathbf{N}(\mathbf{V})_a &= \sum_b \mathbf{N}_{ab} \mathbf{V}(\mathbf{S})_b = \mathbf{0} \quad \text{for } a \neq g, \quad \text{and} \\ \mathbf{N}(\mathbf{V})_g &= \sum_b \mathbf{N}_{gb} \mathbf{V}(\mathbf{S})_b = \mathbf{S} \quad \text{for } a = g. \end{aligned}$$

Hence we need to invert \mathbf{N} to solve for \mathbf{V} . The structure of \mathbf{N} is relatively simple, but there seems to be no (if any) simple inverse. Note that $\mathbf{N} = \mathbf{D} - \mathbf{M}$ is the difference of a (block) diagonal matrix \mathbf{D} and a matrix

\mathbf{M} depending on the μ 's:

$$\begin{aligned}\mathbf{D}_{aa} &= \hat{\boldsymbol{\Omega}}_a^{-1} \otimes \hat{\boldsymbol{\Omega}}_a^{-1}, \\ \mathbf{M}_{ab} &= \mu_{ab} \mathbf{I}_p \otimes \mathbf{I}_p.\end{aligned}$$

In terms of the μ 's we obtain to first order that

$$\mathbf{N}^{-1} = (\mathbf{D} - \mathbf{M})^{-1} \approx \mathbf{D}^{-1} + \mathbf{D}^{-1} \mathbf{M} \mathbf{D}^{-1},$$

yielding the approximation

$$\begin{aligned}\hat{\boldsymbol{\Omega}}_g(\mathbf{S}) &\approx \hat{\boldsymbol{\Omega}}_g + (\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg} \hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{S}) \\ (S12) \quad &= \hat{\boldsymbol{\Omega}}_g + \hat{\boldsymbol{\Omega}}_g \mathbf{S} \hat{\boldsymbol{\Omega}}_g + \mu_{gg} \hat{\boldsymbol{\Omega}}_g^2 \mathbf{S} \hat{\boldsymbol{\Omega}}_g^2,\end{aligned}$$

where $\hat{\boldsymbol{\Omega}}_g = \hat{\boldsymbol{\Omega}}(\mathbf{0})$. To a first order in μ_{gg} this is the same as the approximation

$$\hat{\boldsymbol{\Omega}}_g(\mathbf{S}) \approx \hat{\boldsymbol{\Omega}}_g + (\hat{\boldsymbol{\Omega}}_g^{-1} \otimes \hat{\boldsymbol{\Omega}}_g^{-1} - \mu_{gg} \mathbf{I}_p \otimes \mathbf{I}_p)^{-1}(\mathbf{S}).$$

We also need an approximation for $\ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{S})|$. By first-order Taylor expansion around $\mathbf{S} = \mathbf{0}$ we have

$$\begin{aligned}\ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{S})| &\approx \ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{0})| + \sum_{j,j'} \text{tr} \left[\hat{\boldsymbol{\Omega}}_g^{-1}(\mathbf{0}) \frac{\partial \hat{\boldsymbol{\Omega}}_g}{\partial S_{jj'}} \right] S_{jj'} \\ (S13) \quad &\stackrel{(S12)}{\approx} \ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{0})| + \text{tr} \left[\hat{\boldsymbol{\Omega}}_g^{-1}(\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg} \hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{S}) \right] \\ &= \ln|\hat{\boldsymbol{\Omega}}_g(\mathbf{0})| + \text{tr}(\mathbf{S} \hat{\boldsymbol{\Omega}}_g + \mu_{gg} \hat{\boldsymbol{\Omega}}_g \mathbf{S} \hat{\boldsymbol{\Omega}}_g^2),\end{aligned}$$

where we have used that $\frac{d}{dt} \ln|\mathbf{A}(t)| = \text{tr}[\mathbf{A}(t)^{-1} \frac{d\mathbf{A}}{dt}]$ and $\frac{\partial \hat{\boldsymbol{\Omega}}_g}{\partial S_{jj'}} \approx (\hat{\boldsymbol{\Omega}}_g \otimes \hat{\boldsymbol{\Omega}}_g + \mu_{gg} \hat{\boldsymbol{\Omega}}_g^2 \otimes \hat{\boldsymbol{\Omega}}_g^2)(\mathbf{E}_{jj'})$. We now have the necessary equations to derive the FKL approximate cross-validation score.

Define

$$(S14) \quad f(\mathbf{A}, \mathbf{B}) = \ln|\mathbf{B}| - \text{tr}(\mathbf{B}\mathbf{A})$$

by which the identity

$$(S15) \quad \sum_{i=1}^{n_g} f(\mathbf{S}_{ig}, \boldsymbol{\Omega}_g) = n_g f(\mathbf{S}_g, \boldsymbol{\Omega}_g)$$

holds for all g . The full likelihood (3) in terms of f is given by

$$(S16) \quad \mathcal{L}(\{\boldsymbol{\Omega}_g\}; \{\mathbf{S}_g\}) \propto \sum_{g=1}^G \frac{n_g}{2} \left\{ \ln|\boldsymbol{\Omega}_g| - \text{tr}(\boldsymbol{\Omega}_g \mathbf{S}_g) \right\} = \sum_{g=1}^G \frac{n_g}{2} f(\mathbf{S}_g, \boldsymbol{\Omega}_g),$$

while the likelihood of a single \mathbf{S}_{ig} is

$$(S17) \quad \mathcal{L}_{ig}(\boldsymbol{\Omega}_g; \mathbf{S}_{ig}) \propto \frac{1}{2} \left\{ \ln|\boldsymbol{\Omega}_g| - \text{tr}(\boldsymbol{\Omega}_g \mathbf{S}_{ig}) \right\} = \frac{1}{2} f(\mathbf{S}_{ig}, \boldsymbol{\Omega}_g).$$

In our setting, the fused LOOCV score is given by:

$$\begin{aligned}
\text{LOOCV} &= -\frac{1}{n_{\bullet}} \sum_{g=1}^G \sum_{i=1}^{n_g} \mathcal{L}_{ig}(\hat{\Omega}_g^{-ig}; \mathbf{S}_{ig}) \\
&\stackrel{(S17)}{=} -\frac{1}{n_{\bullet}} \sum_{g=1}^G \sum_{i=1}^{n_g} \frac{1}{2} f(\mathbf{S}_{ig}, \hat{\Omega}_g^{-ig}) \\
&= -\frac{1}{n_{\bullet}} \sum_{g=1}^G \frac{1}{2} \sum_{i=1}^{n_g} \left[f(\mathbf{S}_{ig}, \hat{\Omega}_g) + f(\mathbf{S}_{ig}, \hat{\Omega}_g^{-ig}) - f(\mathbf{S}_{ig}, \hat{\Omega}_g) \right] \\
&\stackrel{(S15)}{=} -\frac{1}{n_{\bullet}} \sum_{g=1}^G \frac{n_g}{2} f(\mathbf{S}_g, \hat{\Omega}_g) - \frac{1}{n_{\bullet}} \sum_{g=1}^G \frac{1}{2} \sum_{i=1}^{n_g} \left[f(\mathbf{S}_{ig}, \hat{\Omega}_g^{-ig}) - f(\mathbf{S}_{ig}, \hat{\Omega}_g) \right] \\
&\stackrel{(S16)}{=} -\frac{1}{n_{\bullet}} \mathcal{L}(\{\hat{\Omega}_g\}; \{\mathbf{S}_g\}) - \frac{1}{2n_{\bullet}} \sum_{g=1}^G \sum_{i=1}^{n_g} \left[f(\mathbf{S}_{ig}, \hat{\Omega}_g^{-ig}) - f(\mathbf{S}_{ig}, \hat{\Omega}_g) \right] \\
&\stackrel{(S14)}{=} -\frac{1}{n_{\bullet}} \mathcal{L}(\{\hat{\Omega}_g\}; \{\mathbf{S}_g\}) - \frac{1}{2n_{\bullet}} \sum_{g=1}^G \sum_{i=1}^{n_g} \left[\ln|\hat{\Omega}_g^{-ig}| - \text{tr}(\hat{\Omega}_g^{-ig} \mathbf{S}_{ig}) - \ln|\hat{\Omega}_g| + \text{tr}(\hat{\Omega}_g \mathbf{S}_{ig}) \right].
\end{aligned}$$

Now, substitution of (S12) and (S13) gives the FKL approximate cross-validation score as an approximation to the fused LOOCV score:

$$\text{LOOCV} \approx \widehat{\text{FKL}} = -\frac{1}{n_{\bullet}} \mathcal{L}(\{\hat{\Omega}_g\}; \{\mathbf{S}_g\}) + \frac{1}{2n_{\bullet}} \sum_{g=1}^G \sum_{i=1}^{n_g} \zeta_{ig},$$

where

$$\begin{aligned}
\zeta_{ig} &= \text{tr}(\hat{\Omega} \mathbf{S} \hat{\Omega} + \mu_{gg} \hat{\Omega}^2 \mathbf{S} \hat{\Omega}^2) - \text{tr}(\mathbf{S} \hat{\Omega} + \mu_{gg} \hat{\Omega} \mathbf{S} \hat{\Omega}^2) \\
&= \text{tr}(\hat{\Omega} \mathbf{S} \hat{\Omega}) + \mu_{gg} \text{tr}(\hat{\Omega}^2 \mathbf{S} \hat{\Omega}^2) - \text{tr}(\mathbf{S} \hat{\Omega}) - \mu_{gg} \text{tr}(\hat{\Omega} \mathbf{S} \hat{\Omega}^2) \\
&= \text{tr}(\mathbf{S} \hat{\Omega}^2) + \mu_{gg} \text{tr}(\mathbf{S} \hat{\Omega}^4) - \text{tr}(\mathbf{S} \hat{\Omega}) - \mu_{gg} \text{tr}(\mathbf{S} \hat{\Omega}^3) \\
&= \text{tr}[\mathbf{S}(\hat{\Omega}^2 - \hat{\Omega})] + \mu_{gg} \text{tr}[\mathbf{S}(\hat{\Omega}^4 - \hat{\Omega}^3)] \\
&= \mathbf{y}_{ig}^{\top} (\hat{\Omega}^2 - \hat{\Omega}) \mathbf{y}_{ig} + \mu_{gg} \mathbf{y}_{ig}^{\top} (\hat{\Omega}^4 - \hat{\Omega}^3) \mathbf{y}_{ig}.
\end{aligned} \tag{S18}$$

To arrive at (S18) we have used the linear and cyclic properties of the trace operator. As $\mathbf{S} = \mathbf{y}_{ig} \mathbf{y}_{ig}^{\top}$, the cyclic property implies the final equality since $\text{tr}(\mathbf{S} \mathbf{A}) = \text{tr}(\mathbf{y}_{ig} \mathbf{y}_{ig}^{\top} \mathbf{A}) = \text{tr}(\mathbf{y}_{ig}^{\top} \mathbf{A} \mathbf{y}_{ig}) = \mathbf{y}_{ig}^{\top} \mathbf{A} \mathbf{y}_{ig}$. Equation (S18) is equivalent to the summand in (S7).